LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES
3 4456 0452276 4

OAK RIDGE NATIONAL LABORATORY
OPERATED BY UNION CARBIDE CORPORATION · FOR THE DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this document, send in name with document and the library will arrange a loan.
UCN-7969
(3 3-67)

ORNL/BCTIC-2

cy. 83

A Review of

INFORMATION PROCESSING IN MEDICAL IMAGING

Proceedings of the Fifth International Conference held at Vanderbilt University, Nashville, Tennessee

JUNE 27 - JULY 1, 1977

NOTE
This work partially supported by the
Bureau of Radiological Health, Food and Drug Administration

BIOMEDICAL COMPUTING TECHNOLOGY INFORMATION CENTER

C-48

Printed in the United States of America. Available from
National Technical Information Service
U. S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
Price: Printed Copy $19.00; Microfiche $3.00

This report was prepared as an account of work partially sponsored by the United States Government. Neither the United States nor its Department of Energy or the FDA Bureau of Radiological Health, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

ORNL/BCTIC-2

Contract No. W-7405-eng-26

BIOMEDICAL COMPUTING TECHNOLOGY INFORMATION CENTER

A REVIEW OF INFORMATION PROCESSING IN MEDICAL IMAGING

Proceedings of the Fifth International Conference
Held at Vanderbilt University, Nashville, Tennessee
June 27 - July 1, 1977

Compiled and edited by

A. Bertrand Brill and Ronald R. Price, Vanderbilt University
William J. McClain and Mildred W. Landay, BCTIC

NOTE:

This work partially supported by the
HEW/FDA Bureau of Radiological Health

Date Published: January 1978

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37830
operated by
Union Carbide Corporation
for the
Department of Energy

LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES

3 4456 0452276 4

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
Price: Printed Copy $19.00 Microfiche $3.00

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor any of its employees, nor any of its contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

CONTENTS

Page

Foreword. vii

Cine-Densitometric Analysis of Coronary Arteriograms
T. Sandor and S. Paulin. 1

A Survey of Computer Processing of Chest Radiographs
R. P. Kruger 13

Processing the Digitised Radiographic Image: Can Decision Aiding Ranges Be Used as an Alternative to ROC Curves
N. J. G. Brown, A. T. Elliott and K. E. Britton. 47

New Methods for Pulsatile Blood Flow Measurement Using Cine-Densitometry
U. Efron, Dan Kedem, Drora Kedem, C. W. Smith, J. Erickson, R. R. Price, J. H. Nelson, and A. B. Brill 66

Topics in Ultrasound Image Processing - A Panel Discussion
Chairman, J. Goddard 85

High-Speed Ultrasonic Scanners
W. A. Anderson 89

Quantitative Imaging from Transmission Ultrasound
J. F. Greenleaf, S. A. Johnson and B. Rajagopalan. 103

Tissue Characterisation by an Analysis of the Distribution of Attenuation Values in C-T Scans
B. R. Pullan, R. Fawcett and I. Isherwood. 116

Fully-Three-Dimensional Image Reconstruction Using Series Expansion Methods
Martin D. Altschuler and Gabor T. Herman 125

Comparison of Radiography and Computed Tomography
D. J. Goodenough, K. E. Weaver and D. O. Davis 143

Three-Dimensional Reconstruction in Longitudinal Tomography by Means of Iterative Approximation
T. Tomitani and E. Tanaka. 174

Positron Emission Computed Tomography Using Large Area Detectors
Frank Atkins, Gerd Muehllehner, and Paul V. Harper 195

Three-Dimensional Reconstruction Based on Sequential Information from a Scintillation Camera
C. Berche, B. Aubert, A. Bethencourt and R. Di Paola 214

Emission Tomography with Time-Coded Apertures
Kenneth F. Koral, Glenn F. Knoll, and W. Leslie Rogers 252

The Exponential Convolution Algorithm for Emission Computer Axial Tomography
O. J. Tretiak and P. Delaney 266

IAEA Image Processing Intercomparison (a discussion)
Chairman, Stephen M. Pizer 279

Page

Computer Display of Organ Surfaces
Gabor T. Herman and Hsun Kao Liu 283

Interactive Biomedical Image Processing
E. L. Hall and R. C. Gonzalez 302

Data Processing and Functional Imaging in Radionuclide Ventriculography
H. Geffers, W. E. Adam, F. Bitter, H. Sigel, H. Kampmann. 322

A Technique for Iterative Estimation of Left Heart Ejection Fractions and Regurgitant Fractions from Noninvasive Radiocardiographic Data
D. B. Twieg 333

Summary of Panel Discussion Regarding the Evaluation of Left Ventricular Function: Gating Versus RCG Technique
Moderator; Dennis Kirch 357

Quantitative Studies with the Gamma-Camera: Correction for Spatial and Energy Distortion
F. Soussaline, A. E. Todd-Pokropek, C. Raynaud. 360

Online Digital Methods for Correction of Spatial and Energy Dependent Distortion of Anger Camera Images
Leonard Shabason, Dennis Kirch, Michael LeFree and William Hendee . 376

Algorithms for EKG-Gated Cardiac Scintigraphy with High Frame Rate and Rapid Turn-Around
Mark L. Moore and Stephen M. Pizer. 389

A Stochastic Method for Determining the Background-to-Liver Ratio in Liver Scanning
B. E. Oppenheim and K. Rezai-Zadeh. 415

Modularity and Command Languages in Medical Computing
A. E. Todd-Pokropek, D. Plummer and S. M. Pizer 426

The Design and Evaluation of a Command Processing System for Scintigraphic Image Analysis
Bruce R. Line, Gerald S. Johnston, and James J. Bailey. 456

Command Languages (a discussion)
Chairman, Stephen M. Pizer. 468

Factor Analysis of Sequence Scintigrams
P. Schmidlin. 472

Identification of Renograms by Factor Analysis
B. E. Oppenheim 481

Carotid and Cerebral Blood Flow
K. E. Britton, C. C. Nimmon, T. Y. Lee, P. H. Jarritt, M. Granowska, A. Greening and Joan M. McAlister 499

Analysis of Errors and Systematic Biases in the Calculation of the Renal Retention Function
F. Erbsmann, J. Struyven, H. Ham, A. Piepsz 526

Page

Validation of Renal Transit-Time Calculations Using Compartmental Models and Direct Measurements
Ronald R. Price, J. J. Touya, R. Branch, J. Goddard, and A. B. Brill 536

Some New SIRT-type Reconstruction Algorithms
A. V. Lakshminarayanan and A. Lent 555

Contrast Enhancement of Scintigrams by the Comparison Wedges Method
I. Runczik, V. Černoch, B. Vavrejn and J. Kidery 584

Comparison of Some Nonlinear Smoothing Methods
P. R. Bell and R. S. Dillon 596

The Implementation of a Floppy Disc Data Logger for Medical Images: A Practical Approach to the Reporting Station Concept
N. J. G. Brown, D. F. R. Burns and K. E. Britton 605

Fluoroscopic Subtraction Imaging Via Real-Time Digital Video Processing--Current Work
Theodore L. Houk, Robert A. Kruger, Charles A. Mistretta, Mitchell M. Goodsitt, Chorn-Gang Shaw, Stephen J. Riederer, David C. Flemming, Joseph F. Sackett, and Andrew B. Crummy. 608

Impact of Large Scale Integration (LSI) on Medical Radionuclide Imaging
Robert O. Smith. 623

Multiple Categories from Medical Images: A Supplement to ROC Analysis
Alan B. Ashare 631

A New Method of Emission Tomography Using Compound Collimation
R. A. Vogel, D. L. Kirch, M. T. LeFree, W. C. Klingensmith, R. S. Trow, P. P. Steele 636

Measurement of Left Ventricular Volume Changes with Computerized Radionuclide Angiography
J. Frischknecht, M. Quinlan, D. Kirch, M. LeFree, R. Vogel, P. Steele 647

FOREWORD

The Fifth International Conference on Information Processing in Medical Imaging was held at Vanderbilt University in Nashville, Tennessee on June 27-July 1, 1977. Like the previous meetings in the series, its objective was to bring together the leaders of research in computer techniques in medical image analysis to share the advances they had made, to constructively criticize each other's work, and to develop directions for further research. An additional objective was to present to the community at large a view of the state of the art in this field by publication of these proceedings.

The participants included medical physicists, mathematicians, computer scientists, engineers, and physicians. The conference emphasized mathematical and computer related matters over strictly clinical matters, though the clinical relevance of the techniques discussed was a prime concern. Within this framework, the subject matter was wide ranging. In addition to topics relating to scintigraphy, which were featured in previous meetings, this conference also focused much attention on the processing of radiological and ultrasonic images, the reconstruction, enhancement and characterization of computerized axial tomography (CAT), made with x-rays (transmission CAT) and gamma rays (emissions CAT), and display systems for 2 and 3-D information.

We wish to thank the Program Committee (Dennis Kirch, Charles Metz, Bernard Oppenheim, Stephen Pizer, Ronald Price, Keith Britton, Robert Di Paolo, Brian Pullan, Peter Schmidlin, and Andrew Todd-Pokropek) for their work in putting this conference together, and the Session Chairmen (Charles Metz, John Goddard, Craig Coulam, Bernard Oppenheim, Stephen Pizer, Thomas Gustafsson, Andrew Todd-Pokropek, Dennis Kirch, Lloyd Knowles, Bruce Line and Ronald Price) for their contributions at the meeting and in the production of the discussion summaries which appear in these proceedings.

We also must thank the several U.S. Government organizations whose support was essential to the planning and conduct of the meetings and to the publication of the proceedings. These include the U.S. Department of Energy (formerly, Energy Research and Development Administration), the FDA Bureau of Radiological Health and the Armed Forces Radiobiology Research Institute. Significant contributions were made by several companies which made it possible to support travel costs for foreign participants at the conference. These included major support from Phillips Medical, El Scint, Inc., and Informatek States, Inc., along with contributions from ORTEC, Inc., Ohio Nuclear, Picker, and Searle Radiographics.

The Biomedical Computing Technology Information Center (BCTIC), through its director, Betty F. Maskewitz, who chaired the conference publications committee, agreed to publish these proceedings as a state-of-the-art review.
The conference organizers, the participants, the financial sponsors, and the BCTIC staff join in the hope that these proceedings will contribute to a broader understanding of the problems, the challenges, and the successes in information processing in medical imaging.

-The Editors and Publisher

CINE-DENSITOMETRIC ANALYSIS OF CORONARY ARTERIOGRAMS

T. Sandor and S. Paulin
Department of Radiology
Harvard Medical School and Beth Israel Hospital
Boston, Massachusetts, U.S.A.

ABSTRACT

A combination of three densitometric analyses are described to assess the hemodynamic significance of obstructive lesions in the coronary arterial tree. These are: (1) The measurement of the degree of stenosis, (2) assessment of the myocardial blush on selected areas in the myocardium, and (3) assessment of washout in individual coronary arteries. The measurements are directly carried out on angiograms recorded on 35mm cine film.

INTRODUCTION

In coronary arteriograms radiopaque substance is injected through a catheter into the coronary arteries, which in turn is exposed to x-rays and either single films in rapid series or cine films of the heart are taken. The information from such films is currently extracted subjectively by the radiologist. Computer-aided densitometry can be used to facilitate the extraction of quantitative information from these films. Such techniques are being developed in our laboratories for assessing the hemodynamic significance of coronary arterial obstructions from 35mm cine angiograms.

The quantitation of these lesions and their consequences is carried out in three steps:

(1) The degree of stenosis is determined.

(2) The myocardial blush in areas distal to the obstructive lesion is measured as a function of time and compared to areas associated to a normal artery.

(3) The turnover rate of the contrast medium in an individual blood vessel (washout-curve) is assessed.

METHOD

Densitometry

The densitometric analysis is based on the Lambert-Beer law[1,2] which describes the attenuation of a narrow monochromatic x-ray beam to an intensity I from an incident intensity I_o when passing through a layer of material with mass-per-unit-area X in the following form:

$$I = I_o \exp\left(-\frac{\mu}{\rho} X\right) \qquad (1)$$

where $\frac{\mu}{\rho}$ is the mass attenuation coefficient. From the above relationship, it follows that $\ln (I_o/I)$ is proportional to X provided μ/ρ is constant, that is:

$$\ln \frac{I_o}{I} = \text{const. } X \qquad (1a)$$

When film is used for image detection, the optical density resulting from the transduced radiation can be expressed as:

$$D = \log_{10} \frac{\text{incident radiant flux}}{\text{transmitted radiant flux}} \qquad (2)$$

Comparing Eqs. (1a) and (2) indicates that:

$$D \sim X \qquad (3)$$

Using this concept, densitometric determination of the amount of the attenuating matter can be performed. For polychromatic x-rays, the relationship (1a) is not strictly correct because the mass attenuation coefficient is a function of the energy of the x-ray photons.

To test the validity of Eq. (3), the following experiment was carried out: radiolucent tubings of known diameter were filled with contrast medium and placed in 20cm of water. Cine angiograms of these blood-vessel phantoms were taken on 35mm film (Kodak Double X) with a Phillips 9" image intensifier and were subjected to densitometric analysis. From the digitized image, one-dimensional optical density distributions were extracted perpendicular to the axis of the blood-vessel phantom. The area enclosed by the distributions of the respective phantom images was computed and plotted against the known geometric cross-sectional area of the tubings. The results are shown in Fig. 1 and indicate good linearity in a range of about 1 to 4 mm "blood vessel" diameter, which is the range of interest in coronary angiography. For filtration of the x-ray equipment 1mm of aluminum was used.

Scanning

In conventional cine densitometry, the projected image of the film is used for densitometric analysis[3,4]. In the present case, densitometry is applied directly to the cine film. Since the small image size necessitates a demagnification between the

object and image spaces of at least seven, a scanning aperture of 25 microns x 25 microns was used when blood-vessel profiles were measured and of 50 microns x 50 microns when specified areas within the heart contour were studied.

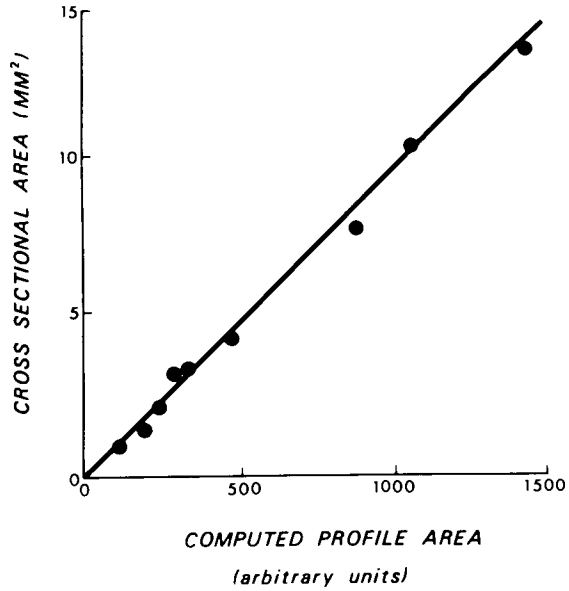


Fig. 1. Measurement of linearity in densitometry by using blood vessel phantoms.

The scanner was a rotating drum-type device that digitized the images into 256 gray levels in an optical density range of 0-2.[5] Out of these 256 levels, about 130 could be distinguished with having statistical significance. The data from the scanner were stored on 9-track magnetic tape, and this tape was used as input for a computer.

Computer and Peripherals

All computations were done by a PDP 11/70 minicomputer located in the Radiology Department of Peter Bent Brigham Hospital. The computer had a core memory of 192K and was interfaced to two RK05 disks of 1.2×10^6 words each and a RP04 disk of 44×10^6 words. Other peripherals included a Tektronix 4014 terminal with a hard copy unit, and a sonic pen digitizer, GRAF-PEN, SAC, Inc., Conn. Several terminals were linked to the computer via dedicated telephone lines; one remote terminal used for this project was a Tektronix 4010 binary graphic terminal in conjunction with the sonic pen digitizer. Both Tektronix terminals had electronic cursors, which could be controlled by thumb wheels. The operating system of the computer was RSX-11D.

Analysis of Stenosis

Densitometric analysis of a stenosis implies a comparison of the averaged amount of contrast medium in the abnormal segment of blood vessel to that in the adjacent normal segment of the same vessel.[6] Formally,

$$\text{Stenosis } (\%) = \left[1 - \frac{\overline{V}a}{\overline{V}n}\right] \times 100 \qquad (4)$$

where $\overline{V}_a$ and $\overline{V}_n$ denote the averaged amount of contrast medium in the abnormal and normal segments, respectively.

The computer program that has been developed to estimate the percentage value of a stenosis according to Eq. (4) has the following main features:

(1) It reads that data from tape to disk and displays them on the Tektronix 4014 storage scope.

(2) The operator may use utility options to facilitate an optimized display of the image such as:

(a) Thresholding: since an angiogram is usually a low contrast image, its display on a storage scope, that has a brightness range of about 10 gray levels, results in an image of very poor quality most of the time. To match the brightness range of the display to the optical density range of the feature of interest on the film, only a selected short range from the density histogram has to be displayed to obtain good quality image. This segmentation or thresholding is a trial-and-error procedure, and the operator may have to repeat it more than once to achieve satisfactory quality.

(b) Magnification: improved visualization of image detail can be facilitated by digital magnification. The operator can request magnification of a rectangular area by identifying its diagonal with electronic cursors.

(3) The part of the picture containing the stenotic artery is identified by an "analyzing window". This is a parallelogram which is defined by the operator by marking three points with the cursors. If not satisfactory, the markings may be repeated. The parallelogram is displayed on the screen with bright solid lines. The operator can use either one window to encompass both the normal and abnormal segments of the blood vessel or one window each for the normal and abnormal segments (Fig. 2).

(4) To reduce the noise in the image, the operator may apply a smoothing filter to the data.

(5) In the next step, the computer extracts a pre-described number of optical-density profiles from the image. These profiles are to be perpendicular to the main axis of the blood vessel.

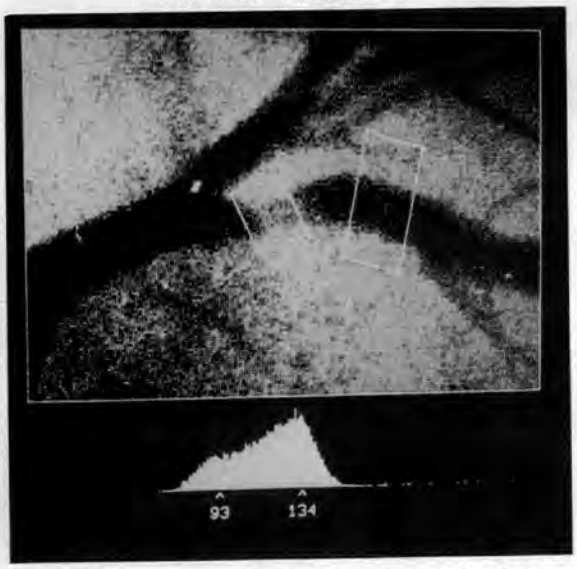


Fig. 2. Analyzing windows encompassing the normal and abnormal segments of a blood vessel.

(6) The operator then marks the boundaries of the profiles with the cursors, and the computer calculates the area enclosed by a straight line connecting the boundary marks and the profile; these data are shown as a bar diagram in the upper right-hand corner of the screen (Fig 3)

(7) After completion of this step, the operator in "conversation" with the computer registers his intentions about how the calculation of the stenosis should take place:

(a) He defines the parts of the bar diagram that

represent the prestenotic, stenotic and post-stenotic segments of the vessel.

(b) He specifies the computation of the percentage value of the stenosis from the pre- or post-stenotic segments or from their mean.

(c) He may repeat the procedure.

(d) He may remark the profiles.

Remote Analysis of Stenosis

The method of analysis described above cannot be used at a location of some distance away from the computer unless the terminal is connected to it by high-frequency cables. To obviate this problem, the following system was developed:

Instead of displaying the image on the storage scope, the operator views the projected image in a film-viewer upon which a sonic pen is mounted. He interacts with this optically-projected image by means of the sonic pen.

To use the projected image for interactive procedures, it is necessary to establish respective coordinate systems for the projected and the scanned images as well as scale factors between them.

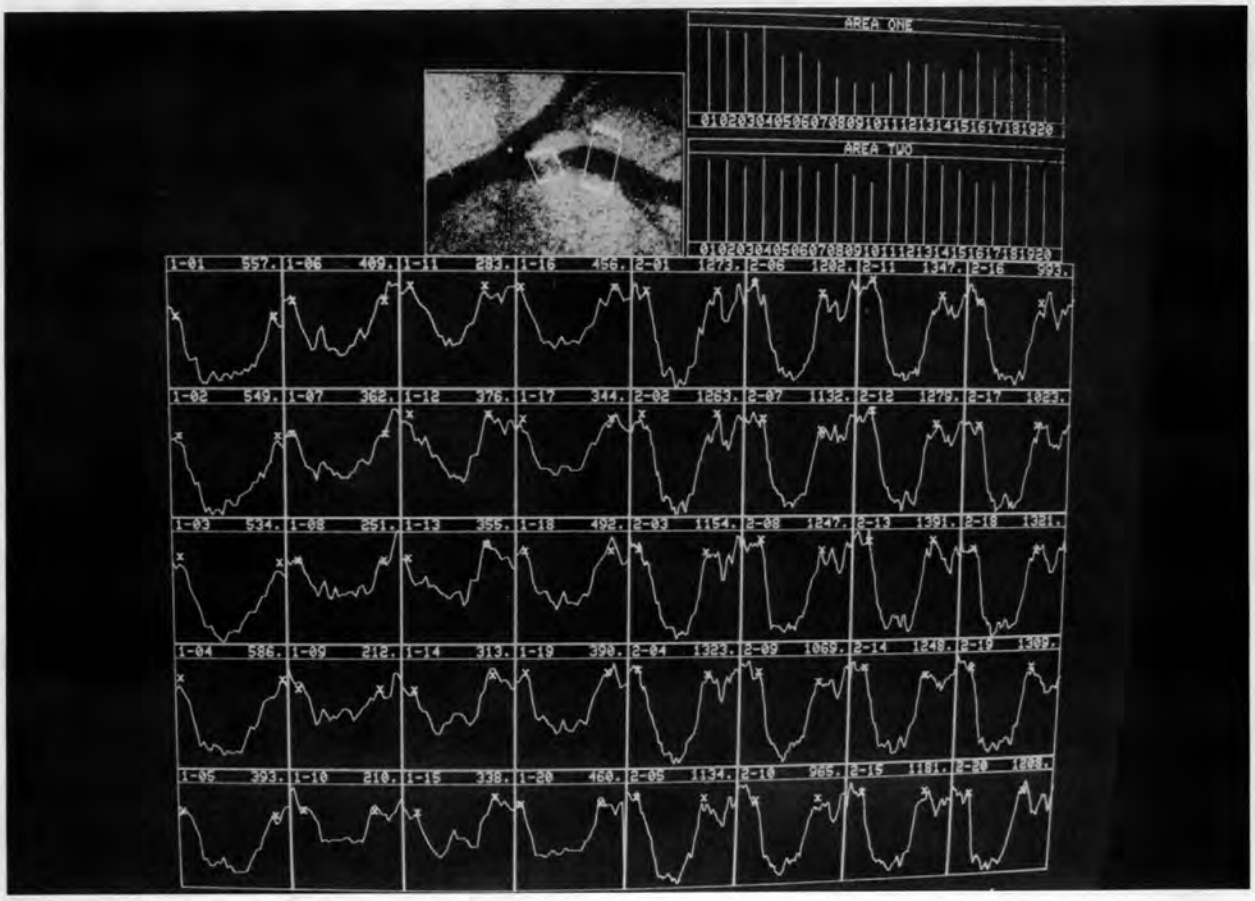


Fig. 3. Densitometric profiles from the abnormal (AREA ONE) and normal (AREA TWO) sections. The bar diagrams indicate the values of area enclosed by the density profiles.

To accomplish this, reference marks on the image have to be selected, which can be easily identified by the operator and found by a computer program. In our system, the boundaries of the cine frame were chosen as reference marks. Identifying positions of the boundaries with the sonic pen provides the necessary information for the computer to link the two coordinate systems together. Identification of the analyzing windows is also performed by the sonic pen. For data feedback, the binary scope (Tektronix 4010) is used. The rest of the analysis proceeds as before. The optical density profiles represent binary information and can, therefore, be sent via regular telephone lines. In our environment, a 4800-baud data transmission rate was achievable. The schematic diagram of the system is shown in Fig. 4.

This system has several advantages:

(a) The high-quality original image can be used for interaction instead of the poor-quality display on the storage scope.

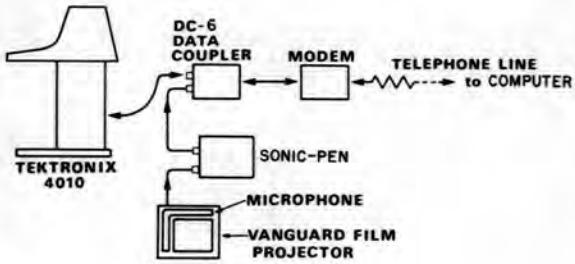
(b) No thresholding is needed.

(c) Several laboratories can share the same computer.

(d) Regular telephone lines can be used for remote terminals.

Reproducibility

The reproducibility of the analysis was tested by three operators, who analyzed stenoses from 20 patients. Cine frames demonstrating stenotic lesions were analyzed by each operator at least three times. From these data, the mean and the standard deviation of the stenosis and the coefficient of variation, commonly defined as the ratio of the standard deviation to the respective mean, were determined for each patient. The results indicated a decrease in the value of the coefficient of variation (variability) with an increasing percent value of stenosis. Figure 5 shows the combined results for the three observers. The least squares fit to the data



arrows indicate data flow

Fig. 4. Schematic diagram of remote analysis of stenotic lesions.

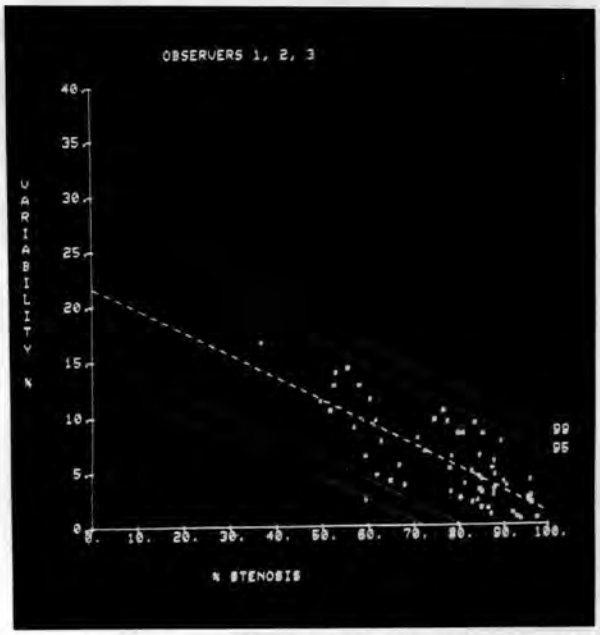


Fig. 5. Reproducibility curve combined for 3 observers.

resulted in a slope of -0.20 ± 0.02. From these data, it followed that the coefficient of variation identified as the intraoperator error is 6% for a 75% stenosis and about 4% for a 90% stenosis.

The interoperator error was also determined by using the above data and the percent stenosis values calculated by the three operators were averaged and the coefficient of variation determined. The results are seen in Figure 6 and the data again indicated a strong dependence of variability on the value of percent stenosis.

Analysis of the Myocardial Blush

Assessment of the myocardial blush implies the determination of the accumulation change of contrast medium in the myocardium with time. Since the shape of the heart varies during the heart cycle, an area identified within the heart image contour will, therefore, change both in shape and size from systole to diastole. To obviate this problem for the analysis, cine frames have to be chosen from the same phase of the heart cycle; this is achieved by using an event marker which is related to the simultaneously-recorded EKG. The basic concept of the procedure is based on a digital subtraction technique and is the following:

The angiographic recording includes a pre- and post-injection part, during which the patient is expected to remain motionless as much as possible. From the pre-injection part, one frame is chosen at a given instant of the heart cycle and this is the "mark", a term borrowed from subtraction radiography. Cine frames for the rest of the analysis are selected from the post-injection part at identical instances of the remaining heart cycles. On each frame the accumulation of the contrast medium is determined by integrating the optical density distribution over specified areas; these values are subtracted from the integral determined for the same area of the mask. Plotting the data against time represents the contrast turnover curve for selected areas in the myocardium.

The computer program which was developed to facilitate the above approach has the following main features:

(1) The cine frames for the analysis are scanned with a scanning aperture of 50 microns x 50 microns and stored on disk.

(2) A constantly well-visualized segment of the border of the cine frame was used to establish a coordinate system automatically for each frame (Fig. 7).

(3) In the next step, the operator chooses a frame from the post-injection part of the study and displays it on the storage scope. He can identify up to three polygons as analyzing windows by using the electronic cursors.

(4) The position coordinates of these polygons are transferred to the rest of the cine frames. The computer carries out the integration and subtraction and plots the data for each window. Since the length of the study covers 7-15 heartbeats, this sets the number of data

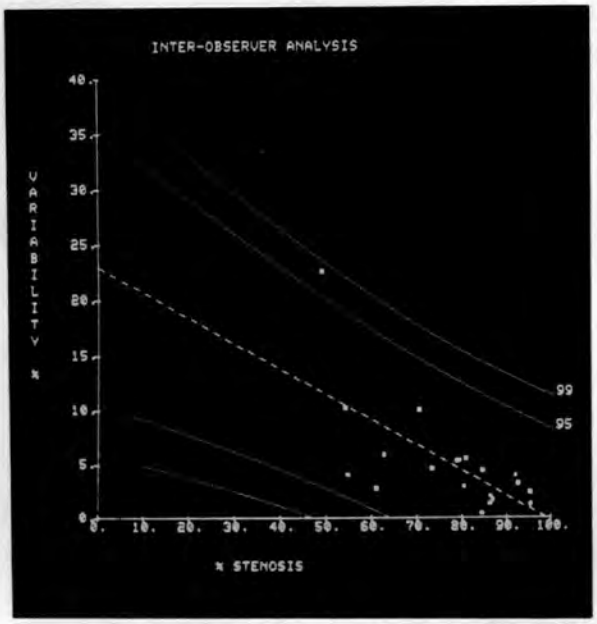


Fig. 6. Analysis of inter-operator error.

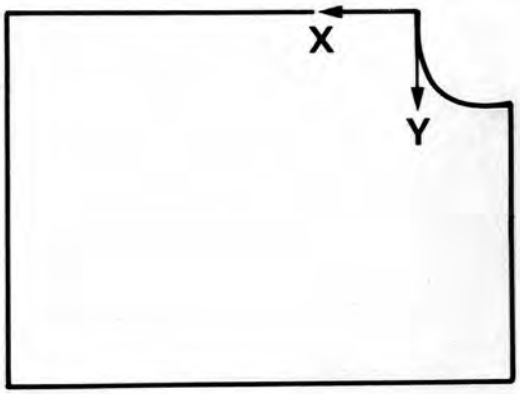


Fig. 7. Cine-frame border with affixed coordinate system.

points available for determining the washout of the myocardial blush. Data points selected for different instances within the heart cycle represent different washout curves. Such difference in washout curves determined for systole and diastole is shown in Fig. 8.

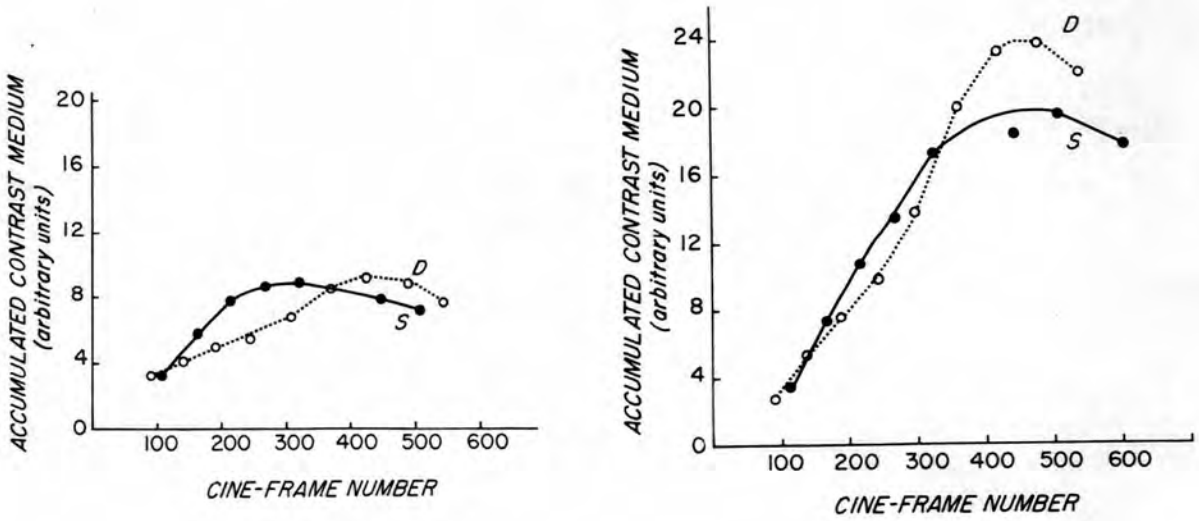


Fig. 8. Examples of myocardial blush curves in systole (S) and in diastole (D) taken at two different areas on the myocardium in a given patient. The length of a regular study was not sufficient to show the full extension of the washout.

Analysis of Washout in Individual Blood Vessels

Propagation of contrast medium within arteries is assumed to reflect grossly the magnitude of existing blood flow. Under normal conditions, simultaneous filling and clearance of the coronary arterial tree is the rule. Differences in washout when observed in individual branches should reflect regional changes in myocardial flow. To measure such changes densitometrically, the variation of the amount of contrast with time within peripheral branches of affected arteries has to be established and compared with unimpaired vascular regions. There are several approaches that we intend to pursue:

(1) The operator identifies on each frame a one-dimensional analyzing window across the vessel image to be analyzed. The optical density profiles will then be extracted by the computer and the corresponding contrast transfer curve will be determined. Such a program has been developed. We found that the poor display capability of our present storage scope is a serious hindrance for this method. The difficulties are most severe at the beginning and at the end part of the washout curve where the image contrast is very low. Repeated thresholding is necessary and this slows down the analysis.

(2) A second alternative is the extension of the principles applied to the analysis of myocardial blush, that is, to use the EKG to select cine frames in identical phases of the heart cycles. This way the analyzing window has to be identified on one frame only and it can be transferred automatically to the rest of the frames. This approach is promising and is currently under development.

(3) Our final aim is to achieve advanced automation of the analysis. To achieve this the transfer of the analyzing window from frame to frame has to be carried out with respect to some anatomic landmark that moves in concert with the selected segment of the blood vessel. A bifurcation of the vessel, for example, might serve for such purpose provided that the visualization of this landmark is adequate throughout the study.

DISCUSSION AND CONCLUSIONS

Visual inspection of cine-radiographic films obtained in conjunction with the selective injection of contrast agents into the coronary arteries of man has been the traditional way of evaluating the degree of coronary artery disease. This analysis is usually performed by observers who are well familiar with the anatomy and patho-physiology of the coronary circulation. Observations of this kind include the scrutiny of the arterial anatomy for the presence or absence of organic changes such as wall irregularities, narrowings or occlusions which may vary in degree, number and location. Additional findings may be the demonstration of compensatory collateral vascular channels as well as variations in the intensity and time delay with which the injected contrast medium is propagated into the different areas of the vascular tree. Although such an interpretation has proven to be of great value in

unequivocal situations of arterial blockage, the validity of such a subjective reading in more complex cases must be questioned, particularly in light of the results of more recent published studies which reveal considerable inter- and intra-observer errors [7,8,9].

The technique developed by us for cine-densitometric analysis of coronary arteriograms is an effort to reduce the above-mentioned uncertainties of subjective interpretation. The described approach is so designed to record objectively the findings to which the observer pays attention. Quantitation of stenoses is accomplished by operator-interactive use of high precision densitometry, which has high reproducibility for narrow lesions. It should be emphasized, however, that the mere identification of a stenosis, even with 100% accuracy, is not sufficient for the characterization of the lesions' hemodynamic impact because of the frequently-developed compensatory collateral pathways of different degree. Our analysis includes, therefore, an effort to analyze changes in contrast propagation and intensity of filling as they occur in the peripheral arterial tree and, consequently, in the intensity of contrast accumulation in the capillary bed, to so-called angiographically-detectable "myocardial blush". This concept has certain similarities with diagnostic techniques of myocardial isotope scanning, where areas of underperfusion are identified by a relatively lower intensity or time delay of activity disappearance. Regional underperfusion of certain areas following selective arterial injections is observed in coronary arteriograms and can be documented densitometrically as shown in our present studies. The physical properties of contrast agents do, however, differ from isotope tracers and it is, therefore, premature to speculate as to what degree the angiographically-demonstrable changes relate to stages of myocardial ischemia or infarction. These questions are being investigated in our ongoing experiments in animal models and compared with the results of other diagnostic methods as used in clinical materials.

Based upon our preliminary results, we are hopeful that the here-described technique will improve information extraction from clinically-performed selective coronary arteriograms.

REFERENCES

1. J. H. Lambert, Photometria sive de Mensura et Gradibus Luminis, Colorum et Umbrae, Augsburg, German, 1870.

2. A. Beer, "Bestimmung der Absorption des roten Lichts in farbigen Flussigkeiten," Ann. Phys. (Leipzig) 86, 78 (1882).

3. H. K. Deininger and V. K. Henck, "Ergebnisse der Kine-densitometric der Nierendurchblutigung", Ro- Bl. 24, 361 (1971).

4. W. Rutishauser, Kreislaufanalyse mittels Rontgendensitometrik, Huber, Bern- Stuttgart Wien, 1969.

5. T. Sandor and G. Congliuso, "Rotating drum scanner-display system for digital image processing," Rev. Sci. Instr. 45, 506 (1974).

6. S. Paulin and T. Sandor, "Densitometric assessment of stenoses in coronary arteries. A pplication of Optical Instrumentation in Medicine IV," Proc. of SPIE 70, 337 (1975).

7. L. M. Zir, S. W. Miller, R. E. Dinsmore, J. P. Gilbert and J. W. Harthorne, "Interobserver variability in coronary angiography," Circulation 53, 627 (1976).

8. B. G. Brown, E. Bolson, M. Frimer and H. T. Dodge, "Quantitative coronary arteriography," Circulation 55. 329 (1977).

9. L. Bjork, H. Spindola-Franco, P. F. Cohn, F. X. VanHouten and D. F. Adams, "A comparison of the diagnostic value of 70mm camera and 16mm cine camera recordings in coronary arteriography," Circulation 50, Suppl. 3, Abst. 292 (1974).

DISCUSSION:

Questions to Dr. T. Sandor focused on technical aspects of the work described.

D. Ochs asked if the images shown in the presentation were taken directly from the Tektronix scope, and, if so, how the gray levels were achieved. Sandor replied that the Tektronix scope was indeed used; and although not marketed as a gray-scale display, it is, in fact, capable of generating about ten distinguishable shades of gray. S. Pizer questioned what was meant by "ten distinguishable shades" and pointed out that the number of such levels can depend upon the pattern in which the levels are presented to an observer, ambient brightness, acceptable error rate, etc. Sandor replied that he had done no observer performance experiments to determine the number of distinguishable gray levels but that he considered his previous statement to be an adequate description of the gray-scale capabilities of the display in question under the circumstances of its use.

B. Pullan inquired whether it was necessary to correct for the variable "gamma" of the image recording system at different levels of exposure. In reply, Sandor stated that no correction is required when using cine film because exposures are well within the linear range of film response. The variation of exposure from one cine film to another is not a problem, since the sensitometric curve of the film used is linear from densities of about 0.2 to 1.8. A serious problem could occur if no corrections were made when using large format X-ray film to image large blood vessels, however, because optical densities behind the vessel could then lie on the "toe" of the film H and D curve.

In response to questions from W. Anderson and C. Metz, Sandor stated that he had not attempted any geometrical corrections for blood vessels that lie in planes not perpendicular to the X-ray beam and agreed that such vessels are difficult to identify from the cine film itself.

A. Todd-Pokropek asked if any attempts had been made to automate the identification of vessel boundaries using edge detection algorithms. Sandor replied that he and his colleagues had tried many such methods but that the results to date have been unreliable and the methods very slow on his computer.

K. Britton inquired whether the quantitative results of this work were used directly by cooperating physicians in the sense that they could establish a percentage stenosis above which they would operate and below which they would not. Sandor explained that at present this was not possible because the decision to operate must be based not only on fractional stenosis but also on other clinical variables. In response to another question from Britton, he stated that simple "meter stick" measurements of vessel diameter directly from the film had resulted in an overall 19% relative error in stenosis determination; whereas the method reported here yielded, for example, about 6% relative error for 75% stenoses.

A. B. Brill and an unidentified member of the audience asked questions concerning the effects of heart cycle phase on the stenosis measurements. Sandor replied that his measurements to date had held phase constant using cine frames keyed on the R-wave of the EKG and that about 14 frames with similar phase were available during a typical study. He stated that he had not yet investigated the extent to which results depended on the phase selected; he doubted much variation but hoped to investigate that point in the future.

A SURVEY OF COMPUTER PROCESSING OF CHEST RADIOGRAPHS*

R. P. Kruger
University of California
Los Alamos Scientific Laboratory
Los Alamos, New Mexico 87545, USA

ABSTRACT

This paper will survey the published work of several investigators who have concerned themselves with various aspects of computer measurement and diagnosis of anatomy present in chest radiography. It is hoped that, at the very least, it will provide a reasonably complete annotated bibliography of this body of research.

INTRODUCTION

During the past decade the field of computer image processing has undergone remarkable growth. From a small nucleus of academic and industrial laboratories, this discipline has expanded into application areas ranging from industrial inspection and earth resources to medicine. Within this latter context there have been numerous modest research efforts directed towards automatic and semiautomatic mensuration and diagnosis of medical radiographs. A particular interest has been shown in the application of scene analysis and pattern recognition algorithms for automated measurement and/or diagnosis of anatomic features recorded on the standard posterior-anterior and/or lateral chest radiograph. The motivation for this research has in part been a recognition that demographically it is the single most prevalent radiographic examination performed. In 1970 approximately 650 million human subject radiographs were exposed. Of this number, 36 percent were chest radiographs.[1]

This survey will be devoted to a historical and chronological review of past and current research and development efforts in this narrowly defined area. The goal is to inform the reader or the perspective researcher of the results of work which have been completed and published. This report will be divided into three major sections: An introduction, followed by scene segmentation algorithms as applied to chest radiography, and finally applications of automated diagnosis.

SCENE SEGMENTATION APPLIED TO THE CHEST RADIOGRAPH

Scene segmentation in its most context-free form is a very difficult and as yet unsolved problem. Given an arbitrary scene, it would necessitate the

---

*Work performed under the auspices of the Energy Research and Development Administration, Contract No. 7405-ENG-36.

division of the scene into a finite recognizable number of components. Two distinct approaches to scene segmentation which have been extensively used can be described by the terms top-down or bottom-up. Top-down scene segmentation refers to a philosophy of parsing a scene into increasingly more specific local components. It relies on a global specification often referred to as a "world model" to guide this search. The bottom-up philosophy by contradistinction initially detects salient local components and uses these with a global model to globally segment the scene. In a completely context-free environment this would require storage of an infinite number of component descriptions using an alphabet of suitable primitive scene descriptors. A paradox exists because in order to detect a scene component, it must be described in some analytic manner. However, a description of the component is needed a priori in order for successful detection to occur. If constraints are applied which limit the number of recognizable scene components but still allow these components to be placed in any orientation and/or spatial position in the scene, the paradox is resolved since some context dependence has been introduced into the world model in the form of a priori knowledge of scene composition. Numerous scene primitive descriptor alphabets have been devised which are orientation and/or position independent. The resulting component descriptions are often robust enough to allow a degree of within-component spatial distortion and still provide for correct description and recognition.

The automated detection and description of major anatomical features in one or several standard radiographic views is in many respects an even more restrictive representation since spatial context dependence of normal anatomy is more or less defined. Consider the standard posterior-anterior (PA) and lateral chest radiographs as stylized in Fig. 1. Both views can be macroscopically divided into a finite number of discrete anatomical components whose locations are interrelated in an approximately fixed anatomical context. Transjuxtaposition of anatomy will not occur, and as a result context is more strictly defined. Scene segmentation algorithms can then be more goal directed and less complex.

Heart and Mediastinal Measurement Algorithms

The work of Becker et al.[2] is the first example of automated feature extraction applied to the PA chest radiograph. This work involved the automated computation of the cardiothoracic ratio (the ratio of heart width to chest width) on 37 PA chest radiographs from the horizontal spatial signature. Kruger[3] expanded this initial work to include the computation of additional indices shown in Fig. 2. Some of the more useful measurements were

1) THr: The thoracic distance is a lateral width of the thoracic cavity at the right diaphragm. This measurement was used to normalize all linear measurements.

2) MDL: The chest midline is a vertical line which divides the thoracic cavity into two more or less equal halves.

3) MLE (MRE): The normalized maximum left (right) extent is the distance from MDL to the left (right) cardiac boundary.

4) CTR: MLE + MRE.

5) TA: The estimated thoracic area is a rectangular approximation of the thoracic cavity and was used to normalize all cardiac projection area measurements.

6) LA(RA): The left (right) cardiac projection area is the normalized area of the heart between the MDL and the left (right) cardiac border.

7) TCA: RA + LA.

8) RP(LP): This is a contour measurement designed to measure the shape of the right (left) vascular band above the right (left) heart apex. The contour is approximated by a fourth-order least-squares polynomial.

The algorithm designed to accomplish these measurements can be best characterized as a bottom-up approach utilizing a collection of heuristic image processing techniques. The goal was to segment the scene by enclosing the cardiac projection in a rectangular box and then creating a binary representation of that projection for later measurement. The technique entailed the use of spatial signatures similar to those mentioned earlier. The initial horizontal and vertical signatures were computed by summing the picture elements in the columns and rows of the image array, respectively. The horizontal signature was very characteristic of the shape noted. Figure 3 illustrates a typical horizontal signature. The MDL was computed first and was chosen as the peak magnitude of the center portion of the horizontal signature. The LEH (left edge of the heart and REH (right edge of the heart) were then chosen as those points where five consecutive first derivatives were either greater than or less than zero. The initial estimate of LEC (left edge of chest) and REC (right edge of chest) were then chosen as secondary signature peaks measured from LEH and REH, respectively. The THr measurement was taken to be the difference between LEC and REC. The TOH (top of heart) and BOH (bottom of heart) were likewise chosen from first derivative analysis of the vertical film profile or signature. The initial estimates of TOH and BOH were then used to create a refined horizontal signature and final estimates of MDL, LEH, REH, LEC, and REC were computed. The image region between the new MDL and LEH was likewise summed to create a refined vertical signature from which TOH and BOH were reestimated. The LEH, REH, TOH, and BOH defined the vertices of the so-called cardiac rectangle as shown in Fig. 3. Within this rectangle an image histogram was computed. This bimodal histogram was then easily thresholded to create a binary image within the rectangle. Once this binary representation was available, it was a relatively straightforward task to complete the heart outline as shown in Fig. 4. This algorithm was applied to 320 chest radiographs on a 256 x 256 pixel grid. This work was an initial example of applied scene segmentation for heart measurement.

This initial work has been refined and improved upon by several researchers. Roellinger[4] used a constrained maximum gradient method to detect the left heart border instead of a thresholding technique. He also refined the estimate of the right cardiac diaphragm intercept by defining it as the coordinate value with shortest distance to the lower right corner of the image.

The output of the cardiac outline program appeared similar to that shown in Fig. 4. At this point Roellinger characterized the shape of the cardiac projection using Fourier descriptors instead of the extent measures and polynomials used by Kruger.[3] TCA was retained from the previous work as a size descriptor. Consider N as the number of boundary points creating a closed perimeter around the cardiac projection. If the boundary points are specified as $(x(n),y(n))$; $n = 0,\ldots,N$, a cumulative angular periodic function $F(n)$ can be defined as the net amount of angular bend between a starting point $n = 0$, and any point $n \leqslant N$ moving clockwise on the closed curve. Note that using this definition $F(0) = 0$ and $F(N) = -2\pi$. Since N will be variable for each projection traced, a new function $G(t)$ was defined over a standard interval for all $0 \leqslant t \leqslant 2\pi$

$$G(t) = F\left(\frac{Nt}{2\pi}\right) + t \ . \tag{1}$$

If G is then transformed by a discrete Fourier transform in polar form, then

$$G(t) = \mu_0 + \sum_m A_m \cos(mt - \alpha_m) \ , \tag{2}$$

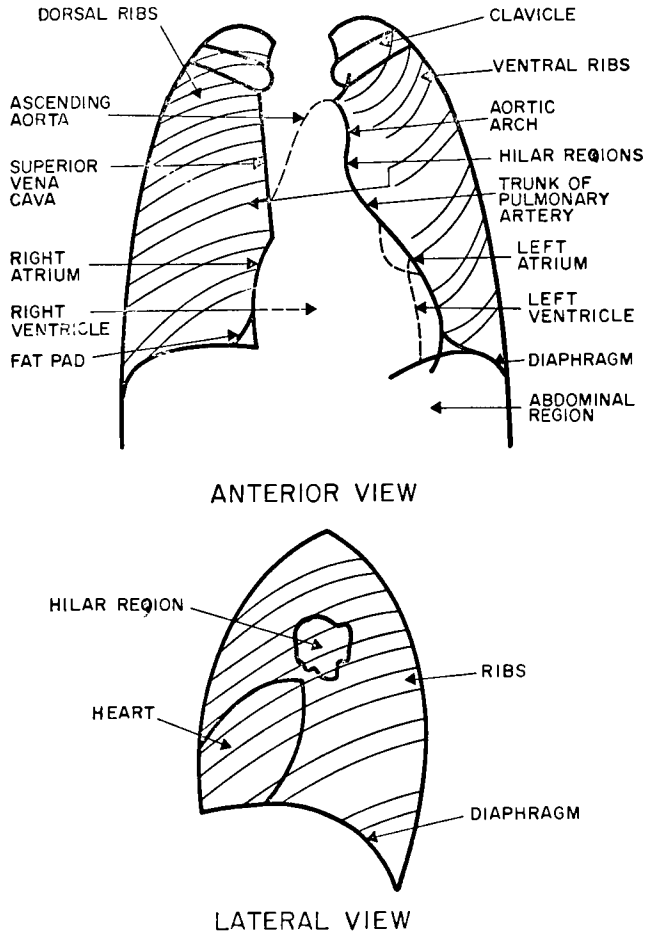


Fig. 1. Labeled outline of major anatomical structure in the PA and lateral chest radiographs.

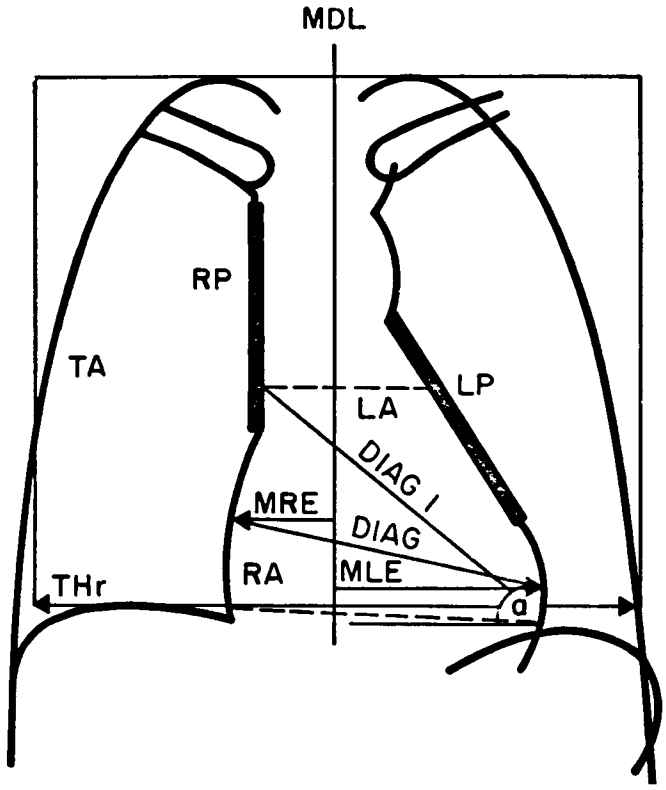


Fig. 2. Horizontal and vertical spatial signatures.

where μ_0 is a d.c. term and (A_m, α_m) are Fourier magnitude and phase, respectively. Thirty pairs of these Fourier descriptors were subsequently used to extract measurements and classify 481 chest radiographs.

Initial software studies led to the development by Sezaki[5] of a vidicon camera minicomputer-based system for rapid (1-second) determination of the CTR measurement. This near-real-time determination used single line scans across the PA film similar in appearance to horizontal signatures. Maximum slope criteria were used to detect the REH and LEH parameters on each scan line. Peak values to the left (right) of REH (LEH) were used to determine the REC (LEC). These four arrays were then used to compute the CTR measurement on 100 chest films.

Lung Outline Detection

The work described up to this point did not outline the lung region *per se* However, this was a logical extension since a good deal of the diaphragm, mediastinum, and outer chest wall had to be detected when heart measurements were made.

Hall[6,7] approached the problem with a bottom-up philosophy using a combination of local binary criteria and global threshold criterion. The global criterion consisted of a threshold computed from a histogram of a 128 x 128 pixel scan of the entire chest radiograph. The bimodal histogram valley was then detected between a low density peak representing the heart, mediastinum, diaphragm, and chest wall and a high density peak representing the lungs. This global threshold philosophy was similar to concurrent work by Spiesberger.[8] The threshold criteria in the former instance consisted of examining the histogram values from a position of one-third the height and to the right of the low density peak (i_p) to a position where three-quarters of the picture elements have been accumulated (i_Q). The threshold (i_T) was selected as the gray level i where initially

$$i_T : h(i-1) < h(i);\ i = i_p, i_Q \ . \tag{3}$$

If the threshold criteria was not satisfied, then i_T was defaulted at

$$i_T = i_p + \frac{(i_Q - i_p)}{2} \ . \tag{4}$$

Next the column sum or horizontal image signature of the low resolution image was obtained. From this the chest midline (MDL) was detected as well as estimates of the left and right lateral chest wall (LEC and REC). This operation divided the chest vertically to isolate the lung regions. The regions to left and right of MDL below the upper quarter of the image (Row 32) were searched for a value above i_T. As one proceeded down the image, the mediastinal and cardiac boundaries were detected in a horizontal region of search defined by [y-6, y+6] where y was the last accepted boundary point. The right and left cardiac diaphragm intercepts (I2, J2)(I5, J5) were found at the minimum distances from the lower right and left image corners respectively. The right hemidiaphragm was found by searching a horizontal distance defined by the array columns from J2 to REC. The vertical region of search began at image row I2 and searched the vertical region [x, 128] where x was the last row value less than i_T. The last accepted coordinate pair defined the right costophrenic coordinates (I3, J3). Identical logic was used to trace the left hemidiaphragm and detect the left costophrenic coordinate (I6, J6). The left lateral chest boundary was found by searching down the image to I6 with a horizontal search region defined by [y-6, 128] for a pixel less than i_T. The right lateral chest

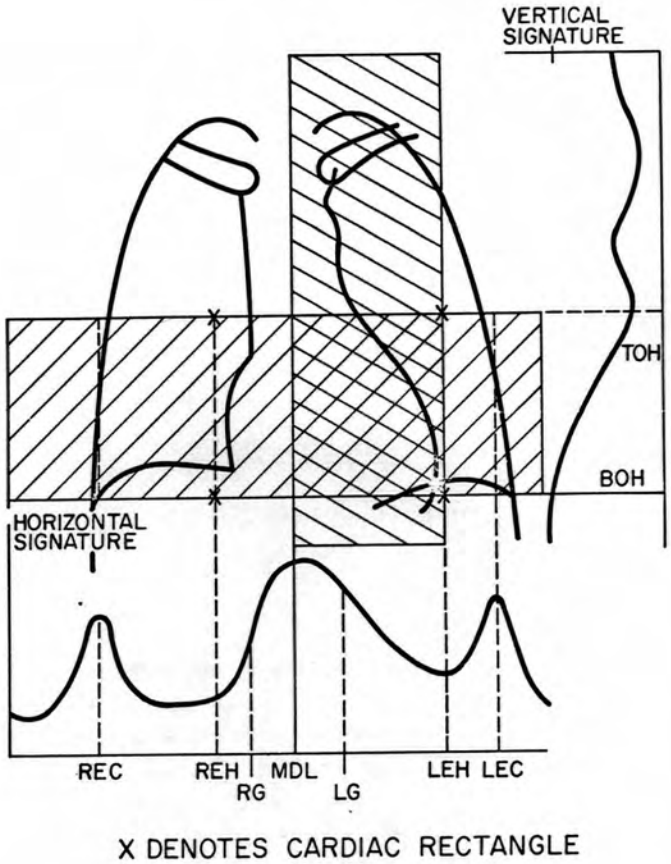


Fig. 3. Anatomical measurements from Ref. [2].

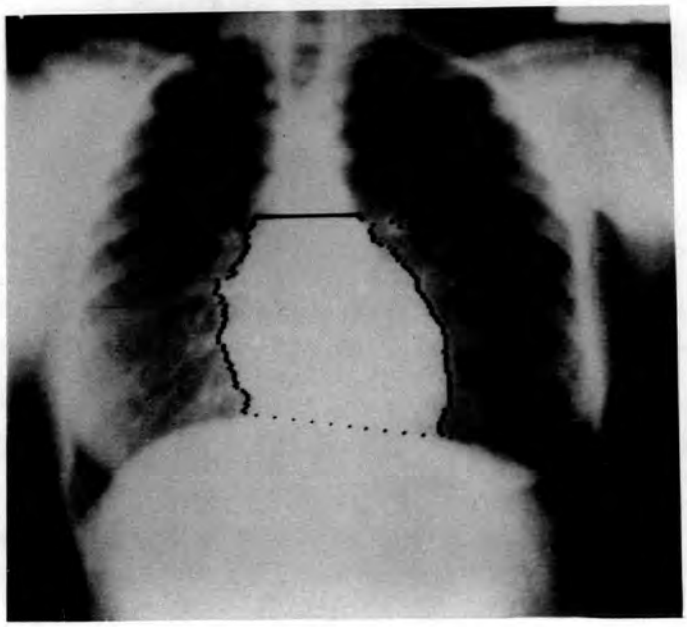


Fig. 4. Completed heart outline.

wall was detected in a similar manner with a vertical region of search from rows 32 to I3 and a horizontal interval of [y + 6,1]. The right or left lung apices (I1, J1), (I4, J4) were chosen as the minimum x coordinate and average y coordinate value associated with these boundary minima.

Each detected boundary was fit to a fourth order least squares polynomial with dual end point constraints. The steps previously described are summarized in Fig. 5. High resolution (512 x 512) digital analysis of just the lung zonal regions was possible because the interpolated polynomial value constituted lateral turn-on and -off indicators as the film was raster-scanned.

This algorithm was performed on 550 chest radiographs using an image dissector camera-minicomputer system as a preprocessing step to computer lung diagnosis. Its weakest feature was the use of a single global histogram threshold criteria. This could lead to local boundary specifications errors. An algorithm with a similar heuristic boundary detection philosophy was concurrently being developed by Tsaing[9] on chest images of similar resolution. The major boundaries are detected using both a local gradient criteria and a global gray scale criterion. One significant preprocessing difference in the latter case was that histogram equalization[10] was performed on the images prior to lung outlining. This operation would destroy the effectiveness of the global criteria chosen by [7,8]. The outlining results of this latter study shown in Fig. 6 on over 400 chest radiographs was impressive. While these latter results were obtained on a large computer, the published computer time per film would leave little doubt of its adaptability to minicomputer-based systems.

Levine[11] broadened the application of lung detection algorithms by extending the technique to also trace the heart and lung boundaries as projected on the lateral view. His motivation was to automate the Barnhard[13] total lung capacity (TLC) calculation. The parameters necessary to accomplish this measurement are shown in Fig. 7. The height (h), width (w), and depth (d) of each segment are measured and the thoracic volume V_t is computed as n truncated elliptical cylinders

$$V_t = \frac{\pi}{4} \sum_{i=1}^{n} w_i d_i h_i \tag{5}$$

where n is the number of sections and h_i the vertical digital scan resolution. The heart volume V_h is calculated as an ellipsoid

$$V_h = \frac{\pi}{6} d_1 d_2 d_3 \quad . \tag{6}$$

The volume V_d of each hemidiaphragm is approximated by one-eighth of an ellipsoid with radius r_1, r_2, and r_3

$$V_d = \frac{1}{8} \left(\frac{\pi}{6} r_1 r_2 r_3\right) . \tag{7}$$

The volume of blood V_B and lung tissue V_{1t} are estimated as computer-stored functions of patient height and weight. The lung volume is then computed as

$$V_L = V_t - V_h - V_d^1 - V_d^2 - V_{1t} - V_B \quad . \tag{8}$$

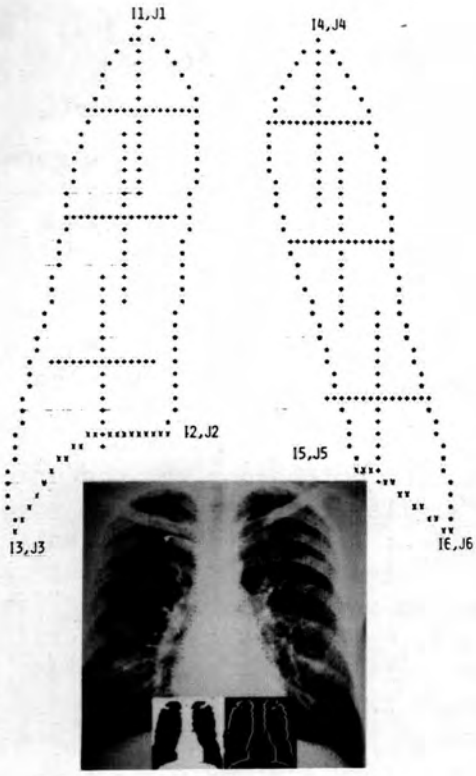


Fig. 5. Lung outline by Hall.

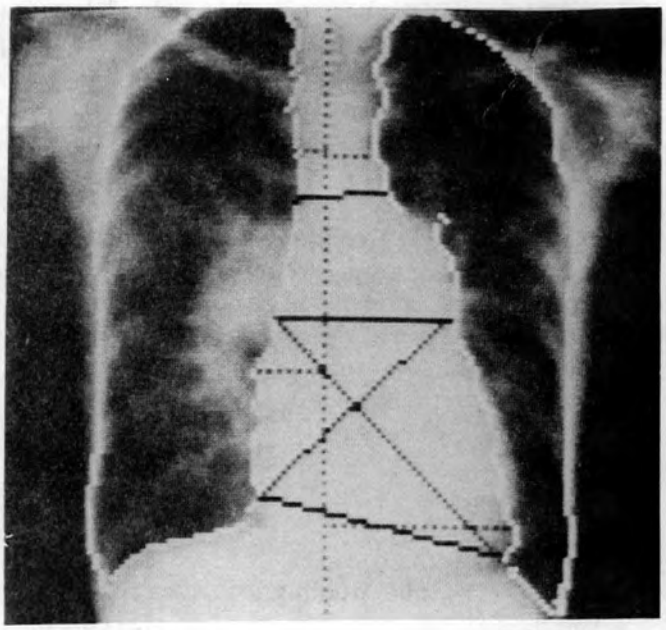


Fig. 6. Lung outline by Tsiang.

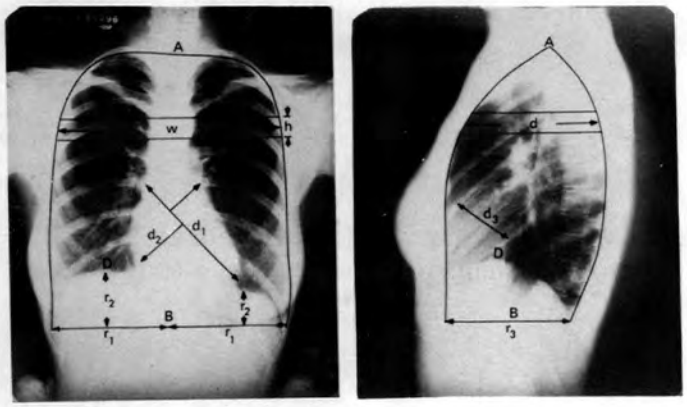


Fig. 7. Barnhard TLC diagram.

The algorithm used for heart detection in the PA view was developed as a refinement of [3]. Once the so-called cardiac rectangle was detected, the border of both the left and right heart boundary was found by means of the Gaussian-weighted first derivative edge detection technique of Ausherman.[12] The use of derivatives in this manner was similar to Tsiang[9] and Roellinger.[4] The diaphragm was detected in a manner similar to [7]. The outer chest wall was detected on both sides by first tracing a short section of lung apex boundary to compute the top of lung (TOL) and lower chest wall. A cubic was then fitted between these traced sections to serve as a guide for detection of the remaining outer chest wall boundary.

A vertical signature of the lateral view was used to detect the region of the diaphragm (BOL) and the top of lung (TOL). The region around this point was searched with an edge detector to compute a top of diaphragm point (TOD). A similarly detected point in the PA view was then used to make the two views colinear. The front edge of chest (FOC) and back edge of chest (BEC) were likewise found from a smoothed horizontal signature similar to those used in [3]. These reference points along with BOL and TOL were used to coarsely segment the scene. Once the lung is enclosed, a rectangular region about the heart in the lateral view is thresholded to create a binary heart projection similar to those presented earlier. This mask is searched for heart-boundary detection. This sequence of operations is illustrated in Fig. 8. At this point the measurements called for in Eqs. (5-8) are undertaken. The results of manual vs automated V_L calculations for 16 PA and lateral pairs indicated a correlation of 0.971 with a standard deviation of 0.262 liters. One of the positive aspects of this work was the direct comparison of manual vs automated V_L in an analytic manner.

The heart and lung outlining algorithms presented up to this point could not be classified as general enough to be extendable into other context dependent scene analysis tasks. While these algorithms may often be superior in implementation efficiency for a specified single task, they will not expand the knowledge of context-dependent scene segmentation as a discipline.

Bailard[14,15] has presented a detailed top-down hierarchic figure-of-merit approach to boundary detection which is applied to both lung and large tumor detection. This technique involves the concept of "planning." The technique attempts to maximize the objective function at image location $z(x_i,y_i) = z_i$.

$$\max_{z_1 \ldots z_n} \left[\sum_{i=1}^{n} G(z_i) + \alpha \sum_{i=1}^{n-2} q\left(z_1, z_{i+1}, z_{i+2}\right) + \beta C\left(z_{n-1}, z_n, z_1, z_2\right) \right] \tag{9}$$

where n is the number of image pixels, $G(z_i)$ is the two-dimensional gradient modulus at the image points z_i; $q(z_i,z_{i+1},z_{i+2})$ is the low-curvature constraint at point z_i and $C(z_{n-1},z_n,z_1,z_2)$ is the curve-length constraint variable.

$$q(z_i,z_{i+ii}) \equiv 1 \text{ if } z_i,z_{i+1},z_{i+2} \text{ subtend an angle equal to } 135^{\circ}, \quad 0 \text{ elsewhere} \tag{10}$$

$$c(\quad) \equiv 1 \; z_n = z_1 \text{ and } z_{n-1}, z_n, z_2 \text{ subtend an angle} \geqslant 135^{\circ} ,$$
$$0 \text{ elsewhere} \quad . \tag{11}$$

The α and β are negative weighting factors. The set of optimal boundary points $z_1 \ldots z_n$ is the set which maximizes Eq. (9). Dynamic programming is used to iterate the solution to a maximum figure of merit. The approach is top-down and hierarchic since the lung boundary is first found at very low resolution. Refinement extends this to higher resolution. At very low resolution such small details as the rib edges and vascular structure are obscured, leading to a complete closed path around each lung. At the next stage the plan-localized search of the higher resolution image is used to refine the initial estimates. Figure 9 shows an example of this lung-boundary extraction procedure. It shows refinement made by the plan followed using Eq. (9), overlaid onto the example radiograph. Application of this technique to large tumor detection will be discussed later. Chien[16,17] used the top-down concept of an embedding metric for the purpose of locating five key anatomical reference points about the right lung on the PA radiograph. These reference points were then connected with straight lines to enclose the lung. These points, shown in Fig. 10, were 1) top of lung (TOL), 2 and 3) intersections of clavicle and the rib-cage projection and mediastinum, respectively, 4) right cardiac diaphragm intercept, and 5) right costophrenic angle.

The five boundary points were found which minimized the cirterion function

$$G(z) = \sum_{i=1}^{n} \sum_{j=1}^{i} g_{ij}(z_i, z_j) \quad , \tag{12}$$

where

$$z = (z_1, \ldots, z_5) \quad .$$

and where g_{ij} is the global evaluation function or cost of embedding the ith component of reference at location z_i; and the jth component at location z_j in the image. In this example, shape information which is a function of the angles θ_{ij} made between the five key points was used as global information. The global criterion was therefore

$$g_{ij}(z_i, z_j) \equiv \left| \theta_{ij} - \tan^{-1}\left(\frac{x_i - x_j}{y_i - y_j} \right) \right| \quad , \qquad i > j \quad . \tag{13}$$

where θ_{ij} was estimated by averaging a priori the angles from several example radiographs. The local criterion was

$$g_{ii}(z_i, z_i) = 1 - C_x{}^i{}_y \qquad i = 1, \ldots, 5 \quad .$$

$C_x{}^i{}_y$ represented the output of a matched filter

$$C_{x_y}^{\;i} = \sum_{\ell}^{M} \sum_{m}^{M} \frac{T^i(\ell, m)\; f\,(x - \ell,\; y - m)}{\| T \| \;\; \| f \|} \tag{14}$$

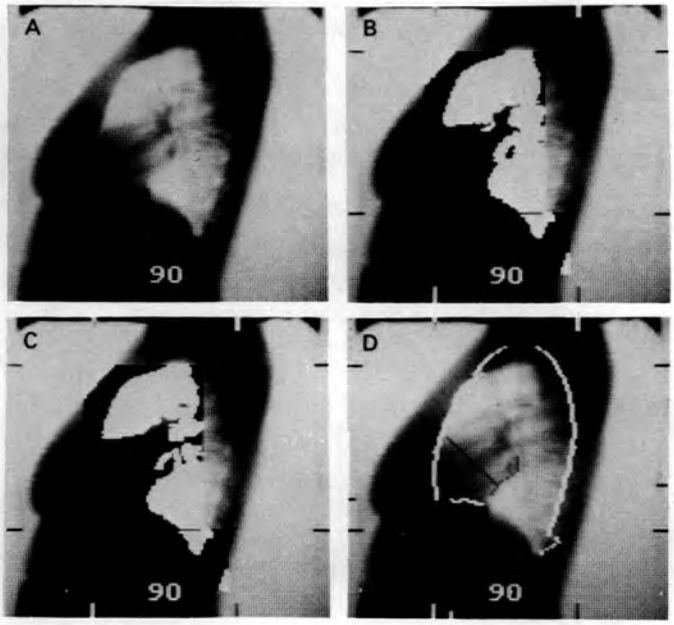


Fig. 8. Laterial view trace from Levine.

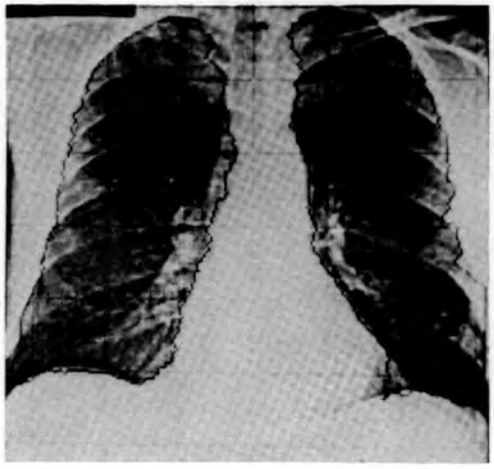


Fig. 9. Lung outline using hierarchic planning algorithm.

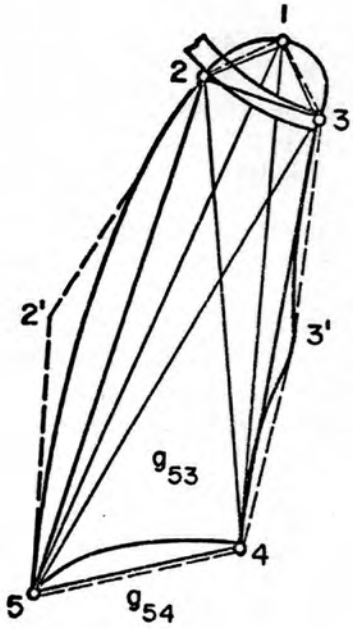


Fig. 10. Graphical outline of embedding metric technique.

where T^i are templates, which were determined a priori, of the local regions surrounding each of the five reference points, and f is the image matrix in that region. At a point of best fit g_{ij} and g_{ii} would all simultaneously approach a minimum that would be zero in the ideal case. Scene context dependence was taken into consideration in that each template T^i pertaining to a particular anatomical reference point was matched only to a prespecified region in f that was selected a priori for the entire population. Figure 11 shows the output of this technique.

The previous two techniques can both be classified as global-to-local or top-down iterative scene analysis techniques with general applicability in any context-dependent class of scenes.

Rib Detection

The detection of the ventral (anterior) and dorsal (posterior) ribs as projected on the PA radiograph are the last macroscopic anatomical structures to be presented. This represents perhaps the most difficult macroscopic detection task in the chest radiograph since their visibility is quite variable.

Toriwaki[18,19] was the first investigator to develop a top-down paradigm for detection of anterior and posterior rib projections as part of the AISCR-V2 (Automatic Interpretation System for Chest Roentgenograms). This approach as recently modified is summarized in the flow chart shown in Fig. 12.

The linear filtering operation consisted of a two-dimensional matched filter with a shape approximately like that of a rib section. Separate matched filters of this type were applied to the inner and outer lung region of each lung. The image output of this operation was then thresholded to create a binary mask or rib images which were traced by an edge-follower. The output of the edge follower for each rib was fitted with a quadratic polynomial which was vertically shifted to serve as an estimate of both the upper and lower rib margins. A similar algorithm was used for both anterior and posterior rib detection.

Two other investigators, Wechsler[20,21] and Persoon[22] have developed algorithms for rib detection. Wechsler[21] has quite recently compared these later two approaches and his comparisons are summarized here.

Both of these later techniques, like Toriwaki, utilized a local rib-edge detector as an integral part of their algorithms. Both the latter approaches utilized contour followers of different philosophy.

Persoon's local edge detector was defined over a 5 x 5 pixel region. As an example, consider the diagram

$$\begin{array}{ccc|cc} f_{11} & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{21} & f_{22} & f_{23} & f_{24} & f_{25} \\ f_{31} & f_{32} & f_{33} & f_{34} & f_{35} \\ f_{41} & f_{42} & f_{43} & f_{44} & f_{45} \\ f_{51} & f_{52} & f_{53} & f_{54} & f_{55} \end{array}$$

The edge detector must decide if f_{33} belongs to an edge and if so it assigns a direction to that edge. As an example the derivative in the right (R) direction

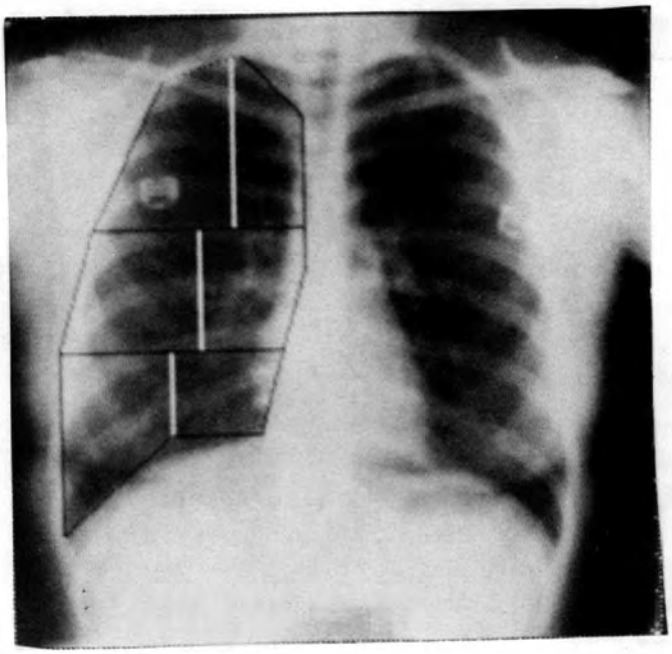


Fig. 11. Piecewise linear approximation of the lung boundary.

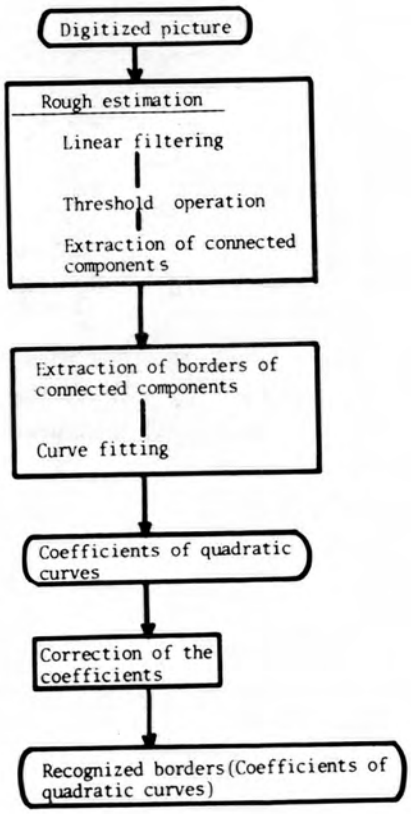


Fig. 12. Toriwaki rib detection method.

is computed. First, pixels f_{i4} and f_{i5}, $i = 1,\ldots,5$, are approximated by a linear model where a_R^* and b_R^* are solved so that

$$\sum_{\substack{i=1,\ldots,5\\ j=4,5}} \left| f_{ij} - a_R^* i - b_R^* \right|^2 =$$

$$\min_{a_R\, b_R} \sum_{\substack{i=1,\ldots,5\\ j=4,5}} \left| f_{ij} - a_R i - b_R \right|^2 \quad . \tag{15}$$

From this model a derivative I_R is also computed

$$I_R = \sum_{\substack{i=1,\ldots,5\\ j=4,5}} \left| f_{ij} - a_R^* i - b_R^* \right| \quad , \tag{16}$$

and a local average gray level A_R defined over the region to the right of the illustrated line is also computed

$$A_R = \frac{1}{10} \sum_{\substack{i=1,\ldots,5\\ j=4,5}} f_{ij} \quad . \tag{17}$$

Similar numbers a_L^*, b_L^*, I_L and A_L are also computed for derivatives to the right. The modified derivative D then becomes

$$D = (A_R - A_L) - (I_R + I_L)/6 \quad . \tag{18}$$

Each 5 x 5 pixel segment is rotated over 360° in eight equal increments with D computed at each rotation. The gradient is the direction for which D is maximized and

$$D_{MAX} = \max_{k=1,\ldots,8} (D_k) \tag{19}$$

is the edge goodness measure. The Persoon contour follower then operates on the image consisting of all $D_{max}(i,j)$ values. Its basic premise is to link together all small edge segments that are nearly horizontal into a convex line which will represent a rib margin. Upper and lower rib margins are compared and quadratic curve fits are again used to extrapolate or fill gaps in order to link edges. This is a bottom-up or local-to-global scene building approach.

Wechsler's edge detector is defined also over a window of size 5 x 5 using a combination of both the gradient and the Laplacian.

The weighted masks shown below are used for approximating the x-derivative, y-derivative and the Laplacian, respectively at point (i,j).

-2	-3		3	2
-3	-8		8	3
-4	-16	i,j	16	4
-3	-8		8	3
-2	-3		3	2

Δx

2	3	4	3	2
3	8	16	8	3
		i,j		
-3	-8	-16	-8	-3
-2	-3	-4	-3	-2

Δy

0	1	1	1	0
1	-2	-1	-2	1
1	-1	0	-1	1
1	-2	-1	-2	1
0	1	1	1	0

L

The outputs of the operator are the quantities V and Q shown as follows:

$$G(i,j) = \sqrt{|\Delta x|^2 + |\Delta y|^2} \quad . \tag{20}$$

The linear combination of G and L forms the operator's magnitude

$$V(i,j) = G(i,j) + 3L(i,j) \quad . \tag{21}$$

The output operator's direction Q is

$$Q(i,j) = k \text{ such that } \frac{k\pi}{2} \leqslant D(i,j) < (k+1)\frac{\pi}{2} \tag{22}$$

$$\text{and } k \in \{1, 2, 3, 4\} \quad ,$$

where $D(i,j) = \tan^{-1}\left(\frac{\Delta y}{\Delta x}\right)$ and $Q(i,j)$ denote quadrants about point (i,j).

Wechsler's rib-margin contour follower is basically of a top-down nature. A Hough array accumulation procedure[14] using V and Q modified for tracing parabolic curves is used to find the lower rib margins from the output Q = 3 of the local edge detector. The upper rib margins are then located using Q = 3,2,1, and 4 respectively. Comparable results of the two algorithms are shown in Fig. 13. The bottom-up approach will be successful if local detectors function accurately. If local information is faulty, it will lack the global guidance necessary to recover. The Wechsler top-down model will perform only as well as the guiding "global world model."

Applications of Computer Diagnosis

The previous sections of this report have summarized in at least minimal detail much of the research concerned with detection and characterization of anatomical macrostructure displayed in the anterior and lateral view chest radiographs. Much of the motivation for this detection and measurement has been the desire to render a machine diagnosis comparable in accuracy to that made by a physician. A machine with the ability to render a consistent normal-abnormal advisory diagnosis with a controlled false positive rate (false abnormal rate) and a "small" false negative (false normal) rate would be of significant medical benefit as a diagnostic tool in mass screening.

The problem of inter- and intra-reader diagnostic inconsistency in chest radiography has been extensively studied by Garland[23] and later by Jerushalmy.[24] The former concluded that experienced film readers are apt to exhibit a false negative rate of 30 percent and a false positive rate of 2% upon a single diagnostic reading. Furthermore, in evaluating pairs of serial radiographs, one experienced physician is apt to disagree with another in a third of the cases.

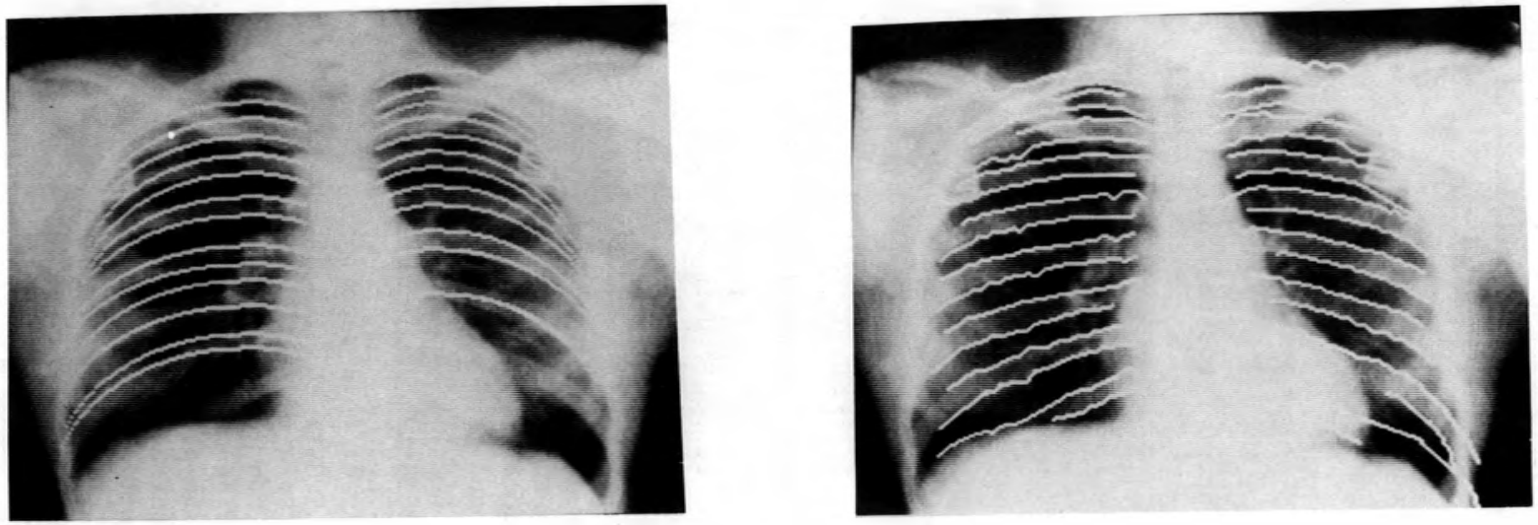


Fig. 13. Comparison of Wechsler's (a) and Persoon's (b) rib detectors.

Upon subsequent review a physician will disagree with himself 20% of the time. These percentages were sustained by the later referenced work on an equivalent number of films and readers. For these reasons, training films for an automated film reader should be validated by other diagnostic tests whenever possible.

Many, but not all, of the machine diagnostic results to be presented in this section utilize some form of Bayesian maximum likelihood decision process with a multi-variate Gaussian class assumption. This represents a form of statistical parametric pattern recognition.

This decision process can be summarized as follows. If $\underline{x} = (x_1,\ldots,x_N)$ represents vector of "selected" features extracted from the image, $[\phi]_{NxN}$ is the intra-class covariance matrix of these features and $\underline{\mu} = (\mu_1,\ldots\mu_N)$ is the intra-class mean vector, the decision function $g_i(\underline{x})$ for each class C_i can be expressed as

$$g_i(\underline{x}) = -\frac{1}{2}\,\underline{x}^t[\phi_i]^{-1}\underline{x} + \underline{x}^t[\phi_i]^{-1}\underline{\mu}_i - \frac{1}{2}\,\underline{\mu}_i\,t[\phi_i]^{-1}\underline{\mu}_i \qquad (23)$$

$$+ \log P(C_i) - \frac{1}{2}\log|\phi_i| \;.$$

In general an unknown $\underline{x}$ is assigned to class C_i when $g_i(\underline{x})$ is maximized. If $[\phi_i]$ is made identical for each class C_i, the classifier is linear in form because the boundaries which separate the diagnostic classes are straight lines in N-space. Unequal $[\phi_i]$ will yield a quadratic N-space boundaries. The so-called modified maximum likelihood decision function is similar in form except that the last two terms of Eq. 23 are deleted.

The testing results to be presented are often the results of a process commonly referred to as "jackknifing." A jackknife is a process whereby z percent of the N-vectors $\underline{x}$ are removed from the entire population of N-vectors $\underline{x}$ prior to classifier training. The z percentage commonly varies between 1 and 50 percent dependency on the size of the sample population. After removal the classifier is trained on the remaining 100-z percent of the data. Finally, the z removed unknowns are submitted and classified. This process is repeated until each vector in the population has been removed and tested. This represents an unbiased testing process for the population.

Classification results to be presented will be restricted to two-class normal-abnormal confusion matrices similar in form to the one shown in Table I even though the specific study may have included several abnormal classes.

Table I. General Form of Two-Class Confusion Matrices

		Individual Reader	
		N	A
True Class	N	True Negative	False Positive
	A	False Negative	True Positive

This is both an attempt at brevity and a means of emphasizing the importance of the normal-abnormal decision in screening applications. In these matrices the numbers in parentheses will represent the actual number of samples classified with the corresponding percentage shown above it.

Film Exposure Classification

Common logic would seem to dictate that films which are improperly exposed will more often be read incorrectly. Jacobson[25] concluded that, in lung diagnosis, overexposed chest films consistently exhibit a higher than expected false-negative percentage, while underexposed films exhibit a higher than expected false-positive percentage. It would therefore be logical for a diagnostic machine to remove improperly exposed films as its initial test.

Kruger[26] has demonstrated that such an exposure prescreening is possible utilizing features available from a gray level histogram of the image. The histogram measurements computed were 1) maximum value, 2) minimum value, 3) pixel count at the histogram's largest peak, 4) mean value, and 5) standard deviation. These five features were classified using a (linear) Gaussian classifier. The two-class normal-abnormal classifier training results using 89 films for device training was 94 percent. Classifier testing results against 96 unknowns yielded 87 percent correct.

Computer Heart Diagnosis

The heart and major vessels above the heart have been computer-classified more than any other anatomical feature in the chest radiograph. Kruger[3] classified the heart and major mediastinal vessels, using the area and extent measures detailed earlier, into a normal and three abnormal classes. The abnormal film data was obtained from catheter, surgery, or autopsy verified rheumatic heart disease patients. Computer testing results using a jacknife of z = 10 percent and a piecewise linear Gaussian classifier are shown in Table II.

Table II. Rheumatic Heart Study (Computer)

		Computer Class N	Computer Class A	
True Class	N	88.6 (78)	11.4 (10)	Percent Correct 92.1
	A	6.2 (12)	93.8 (181)	

In this same study 10 radiologists each read on an average approximately half the 281 films for the same purpose. Their overall group performance was 89.8 percent correct.

In a similar study Tsiang[9] combined area and extent measures similar to those used by Kruger[3] and the Fourier descriptors proposed by Roellinger[4] to computer-classify normal and several congenital heart disease classes. A modified maximum-likelihood Gaussian classifier with a z = 10 percent jackknife yielded the following diagnostic testing results. Verification of data as to class was again by other than radiographic means. Results are shown in Table III.

In this study there was no physician-comparison made. However, the young age of most of the patients and the number of categories of congenital diseases would make this a more difficult diagnosis both for a machine and for a trained film reader.

Table III. Testing Results
Normal vs Congenital Heart Disease

		Computer Class N	Computer Class A	
True Class	N	76.1 (121)	23.9 (38)	Percent Correct 79.4
	A	18.3 (42)	81.7 (187)	

Computer Diagnosis of Pulmonary Disease

Diseases of the lung as recorded on the plain chest radiograph can be grossly divided into the two major categories of nodular and diffuse (disseminated). With nodular disease, there are usually one or a small, and thus countable, number of 1-cm or larger opacities (lesions) visible on the radiograph. These opacities may be indicative of a variety of etiologies. Diffuse lung disease can also have many etiologies (over 200). Radiologic signs of diffuse disease as viewed on the chest radiograph have been broadly classified into the minor categories of alveolar and interstitial disease. Additional classification of the many small 2-mm average size interstitial opacities into the sub-categories of nodular and reticular has allowed a further division of the radiologic signs. The many small lesions representative of disseminated alveolar or interstital disease present a visual opacity pattern that is textural in appearance. Similarly, normal pulmonary vascularity can be characterized in a textural manner. It is, therefore, not unexpected that the initial research on computer diagnosis of disseminated lung disease concentrated on the quantification of these textural differences. These measurements have not attempted to count opacities per se but rather have been designed to provide an indirect quantitative description of their textures.

Optical Textural Measures

The study of computer detection and classification of lung disease through the use of textural discriminators has been accomplished through both optical and digital means. The optical approach has centered about the use of measurements extracted from the Fraunhofer diffraction pattern.

The Fraunhofer diffraction pattern or optical Fourier transform is simply the pattern which results when collimated laser light is used to illuminate a film region f. The resulting transmitted light from that region then passes through a thin convex lens. If α_1 is the focal length of that lens, the optical Fourier transform $\hat{F}$ is reconstructed at a distance α_1. This sequence of operations is shown in Fig. 14. In mathematical form

$$\hat{F}(u,v) = C \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-j2\pi(\pi x + vy)\right] f(x,y)\,dxdy \tag{23}$$

where

$$u = v_x \lambda \alpha_1 \qquad v = v_y \lambda \alpha_1$$

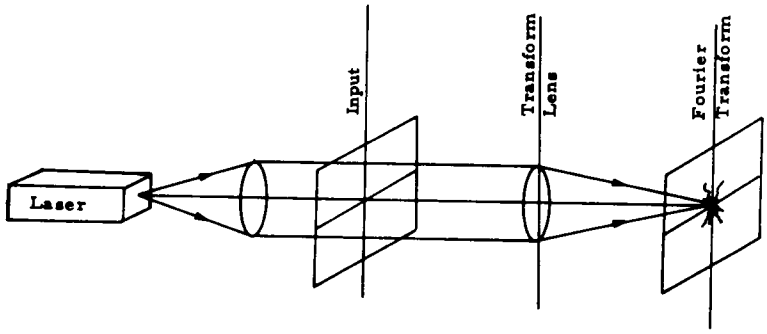


Fig. 14. Optical Fourier transform.

and where v_x and v_y are spatial frequencies, and λ is the wavelength of the coherent source. The modulus or magnitude of the Fourier transform is defined as

$$I(u,v) = \left(F(u,v)\hat{F}^{*}(u,v) \right)^{1/2} , \tag{24}$$

where * denotes the complex conjugate. It is also possible to reverse this process by adding a second transform lens at distance α_1 beyond the transform plane. Frequency domain filtering is easily accomplished by adding a photo mask at α_1.

Becker[27] reported the first use of Fourier domain filtering in chest radiography. His application demonstrated the ability of a simple Fourier-domain wedge filter to reduce the visibility of dorsal rib structure. Revesz[28] performed an interesting experiment using human observers. Three-hundred posterior intercostal (inter-rib) radiographs were submitted to an optical system which recorded the magnitude of the optical Fourier Transform I on film. The films were selected from normal as well as nodular and reticular interstitial disease categories. Six trained observers were then asked to independently classify both the radiographs and their diffraction patterns into these three diagnostic categories. The results indicated that 78 percent of the radiographs were classified correctly. However, 98 percent of the diffraction patterns were classified correctly. Thus, it appeared that there was a characteristic visual appearance of the diffraction pattern which aided visual discrimination.

Additional concurrent work by Biagini[29] demonstrated that the radial signature r_j of the Fourier Modulus I could usually discriminate between normal local lung regions and those which displayed evidence of pulmonary emphysema. In the former example it was the addition of numerous small opacities characteristic of interstitial disease patterns onto normal vascularity which gave rise to additional visually detectable high frequency activity in the Fourier spectra. In the latter case, it was the reduction of this high-frequency energy below that exhibited by normal pulmonary vascularity which reflected the vascular destruction caused by emphysema.

Kruger[30] replaced the photographic recording of the diffraction pattern with a solid state detector designed to make direct digital annular-ring and angular-wedge signature measurements of the optical Fourier transform modulus. If the Fourier modulus is refined in polar coordinates $I(\rho,\theta)$, the ring signature measurements r_j can be defined as

$$r_j = \int_0^{\pi} \int_{\rho_j}^{\rho_j+\Delta\rho} I(\rho,\theta)\rho d\rho d\theta \tag{25}$$

$$j = 1,\ldots,32 \quad .$$

The wedge signature values can similarly be described by

$$w_j = \int_{\rho_{min}}^{\rho_{max}} \int_{\theta_j}^{\theta_j+\Delta\theta_j} I(\rho,\theta)\rho d\rho d\theta \tag{26}$$

$$j = 1,\ldots,32 \quad .$$

The r_j measures are sensitive to differences in spectral energy as a function of radial frequency. The w_j measures are sensitive to differences in angular spectral energy content and are useful for differentiating reticular structure.

These 64 measures were subsequently normalized to unit energy to compensate for linear interfilm differences in dynamic range and mean film density. This apparatus was used to collect data from 6.35-cm-diameter lung regions on 141 zonal chest radiographs. Each zone was roughly 1/3 of the vertical height of the lung from the diaphragm to the apex. There were 39 normal films and 102 abnormal films. The abnormal films were selected from a representative population of individuals with pneumoconiosis, an occupationally related disease associated with the long-term inhalation of coal dust. This malady is classified as a diffuse interstitial lung disease visualized by the superposition of numerous small rounded or irregular opacities onto normal vascular lung patterns. In the absence of a work history containing long-term dust exposure, the abnormal film looks not unlike other diffuse interstitial diseases. A quadratic Gaussian classifier and a jacknife of size z = 50 percent was used to classify this data. The results are shown in Table IV. An example of a normal and advanced pneumoconiosis film is shown in Fig. 15.

Table IV. Two-Class Testing
Diagnosis of Pneumoconiosis

		Computer Class N	Computer Class A	
True Class	N	69.7 (27)	30.3 (12)	Percent Correct 88.7
	A	4.0 (14)	96.0 (98)	

Six specialists in pneumoconiosis diagnosis were asked to read the identical film zones. Their combined results yielded 91.5 percent correct.

Stark[31] replicated and refined this experiment on 64 lung zonal films. Twenty-seven films were normal and 37 were abnormal pneumoconiosis films. Notable changes and refinements were as follows. First, he placed the film in a phase matching liquid gate to reduce spurious diffractions from emulsion relief. The solid state detector was replaced by a vidicon and the r_j and w_j were computed from a digitally stored matrix of the diffraction pattern. A distribution-free k nearest-neighbor classification technique was used instead of a Gaussian classifier. His jacknife test procedure consisted of a one-film-at-a-time removal scheme. The results indicated a correct rate of 90.7 percent. The use of the phase-matching liquid gate represents a refinement which should be noted by future investigators.

The most ambitious project to date has been the construction of a prototype film reading device for the detection of pneumoconiosis in the chest radiograph.[26] This system consists of a digital minicomputer used for both lung detection, and later for diagnosis; a diffraction pattern sampling unit for r_j and w_j feature measurement; and a computer-controlled x-y film transport. The latter two elements were both mounted on an optical bench. A 35- by 43-cm transverse film movement was possible. Previous studies with an optical computer attempted to diagnose only on selected zones of the lung. This device was designed to automatically position the film into six lung zones and integrate these zonal decisions into an overall film decision.

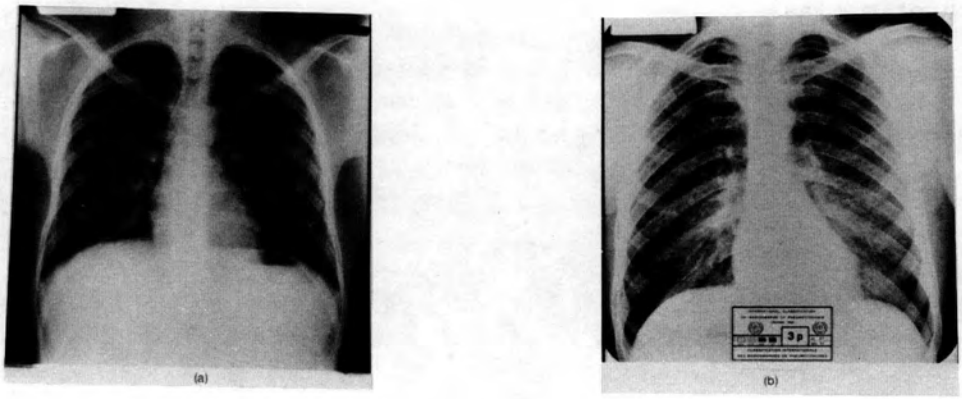


Fig. 15. Normal and Pneumoconiosis film.

An example of a typical lung outline is shown in Fig. 5. The crosses represent the center and extent (7.62-cm diameter) of the six lung regions examined. In this experiment the film library consisted of 465 chest radiographs read independently by 6 highly proficient physicians on both a per zone and per film basis. Because of reader inconsistency, the device was trained to render only a normal or abnormal decision on the 465-film set from training data gathered from a subset of 130 films. Each of the 130 training films had been independently read by the six physicians and judged by each of them either completely normal or abnormal in all six lung zones. There were 57 normal and 73 abnormal films in this training set.

Since there was no discernible pattern for physician integration of individual zonal decisions into an overall film decision, a statistical approach was taken to accomplish this task.

$p(\underline{x},C_i)$ represents an N-dimensional intra-class probability density function. Using the mixed form of Bayes rule, it is also possible to compute $P(C_i/\underline{x})$, the *a posteriori* probability of class membership, C_i, given an $\underline{x}$ for each of m classes.

$$P(C_i/\underline{x}) = \frac{p(\underline{x},C_i)}{p(\underline{x})} \qquad i = 1,\ldots,m \quad , \tag{27}$$

where

$$p(\underline{x}) = \sum_{i=1}^{m} p(\underline{x},C_i) \quad .$$

Thus, choosing the maximum $P(C_i/\underline{x})$ is equivalant to choosing the maximum $g_i(\underline{x})$. In either case one has made a maximum *a posteriori* decision. The use of these *a posteriori* probabilities allows not only the determination of class membership, but also the probabilistic certainty of that decision.

Each zone was trained with a <u>zone specific</u> N = 10 dimension feature vector $\underline{x}_z$ using a quadratic classifier, and the *a posteriori* probabilities were computed. It was decided to produce an overall film decision by summing the *a posteriori* probabilities of normal $P_z(C_n/\underline{x}_z)$ for the six zones as shown in Eq. (28) and thresholding the result.

$$\bar{P}(C_n/\underline{x}) = \frac{1}{6} \sum_{z=1}^{6} P_z(C_n/\underline{x}_z) \quad , \tag{28}$$

where

$$P(\underline{x}_z) = \frac{1}{6} \quad \text{and} \quad \underline{x} = (\underline{x}_1,\underline{x}_2,\ldots\underline{x}_6)$$

and $\underline{x}_z$ is the zone measurement vector and $\underline{x}$ is a vector of all measurements. If $P > T$ then the film is assigned to the normal class.

Furthermore, if it is assumed that $\underline{x}_z$ are 6 unlabeled samples drawn independently from a normal-abnormal mixture distribution, the $\bar{P}$ still constitutes a maximum likelihood decision for the film.

Using this method and a threshold of T = .483, 91.5 percent of the 130 films were classified correctly in training. Diagnostic testing results

on all 465 films with a T = .56 yielded the results of Table V. In order to obtain testing results, a normal-abnormal consensus of 3 out of 5 physicians was used. Neither the computer or the least-consistent physician was used in this consensus.

Table V. Hybrid Computer Results

Consensus Refined Truth		N	A
	N	60.8 (62)	39.2 (40)
	A	16.5 (60)	83.5 (303)

Percent Correct 78.5

This is a severe test since only 1/2 the normal and 1/5 of abnormal films were used in device training.

The averaged physician reader-performance on the same films was 91.6 percent correct.

Goodenough[32] compared the performance of the general population of physicians who were not trained to read radiographs for presence of pneumoconiosis (A readers) with physicians trained in this diagnosis (B readers). He used receiver-operation-curves (ROC) commonly applied in radar detection problems. The resulting curves labeled A and B in Fig. 16 refer to the performance of A and B level readers when compared to truth. A third curve refers to the performance of the hybrid system. This curve was traced by varying T between 0 and 1 for the 465-film test set used in the study. A fourth curve represents the reading level subsequently established by the National Institute for Occupational Safety and Health Administration for certification of all future B-level readers. ROC curves are superior to confusion matrices for displaying the results of reader performance because of their dynamic nature.

Recent tests of the same prototype by Strand[33] have yielded a two-class 82% correct testing result with a one-film-at-a-time-removal jacknife on 111 films. The films were selected from normal and prepulmonary edema categories.

Digital Textural Measures

Sutton[34] first applied the concept of textural discrimination to detection of interstitial fibrosis, a reticular appearing disease, in the chest radiograph. The upper-right lung zone of each radiograph was divided into four quadrants k moving clockwise from the upper-outer quadrant. The measurement calculated in each quadrant was a horizontal first difference

$$E_k = \sum_{\substack{i \\ \text{region } k}} \sum_{j} \left| f(i,j) - f(i+1,j) \right| , \quad k = 1,2,3,4 \ . \tag{29}$$

It was found that $M = E_3/E_4$ was a single feature which, in a normal-abnormal test, classified 20 out of 24 films successfully. The testing procedure was a one-at-a-time jacknife on 9 normal and 15 abnormal films.

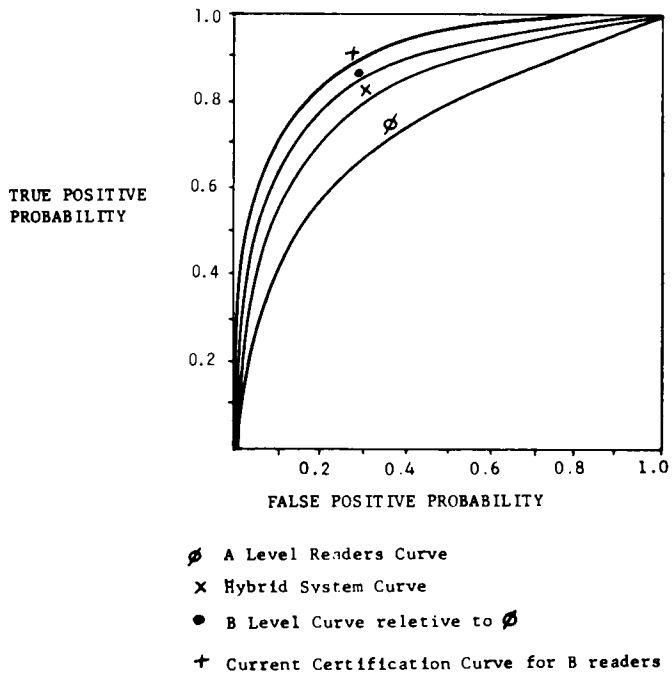


Fig. 16. ROC curves of reader groups.

Kruger[30] next applied the digital texture measures to the detection of pneumoconiosis. It was assumed that this texture-context information was adequately specified by the symmetric matrix of relative frequencies p(i,j) with which two neighboring pixels are separated by a distance (d) and an angle (a) for each (i,j) gray-level pair in the space. A symmetric count matrix was formed within each square with all count matrices summed and normalized to create a matrix of relative frequencies for each interrib space as a function of a and d

$$p(i,j,a,d) \quad , \tag{30}$$

where

$$i = 0,1,\ldots,7, \qquad j = 0,1,\ldots,7,$$

$$a = 0°,45°,90°,135° \qquad d = 1,3,7,11.$$

The following five textural measurements $T_k(a,d), d = 1,\ldots,5$ were computed for each matrix

$$T_1(a,d) = \sum_{i=0}^{7} \sum_{j=0}^{7} ijp(i,j,a,d) \tag{31}$$

$$T_2(a,d) = \sum_{i=0}^{7} \sum_{j=0}^{7} (i - j)^2 p(i,j,a,d) \tag{32}$$

$$T_3(a,d) = \sum_{i=0}^{7} \sum_{j=0}^{7} \frac{p(i,j,a,d)}{1 + (i - j)^2} \tag{33}$$

$$T_4(a,d) = \sum_{i=0}^{7} \sum_{j=0}^{7} p(i,j,a,d) \log p(i,j,a,d) \tag{34}$$

$$T_5(a,d) = \sum_{i=0}^{7} \sum_{j=0}^{7} |i - j| p(i,j,a,d) \tag{35}$$

for a = 0°, 45°, 90°, 135°, and d = 1,3,7,11. T_1 is an autocorrelation measure designed to measure image coarseness. T_2 is a dissimilarity measure often called the moment of inertia. T_3 measures the extent to which the same or similar gray levels tend to be neighbors. T_4 is a conditional entropy measure and measures image homogeneity. T_5 is another dissimilarity measure which is similar to T_2. The number of textural final features was reduced by calculating the mean $\bar{M}$, variance V, and range R at a given distance d for each of 4 angles a.

$$\bar{M}_k(d) = \frac{1}{4} \sum_{a=1}^{4} T_k(a,d) \tag{36}$$

$$R_k(d) = \max T_k(a,d) - \min T_k(a,d) \tag{37}$$

$$V_k(d) = \frac{1}{4} \sum_{a=1}^{4} (T_k(a,d) - \bar{M}_k(d))^2 \tag{38}$$

for k = 1,...5, d = 1,3,7,11, and a = 0°, 45°, 90°, 135°. The final $\bar{M}$, R, and V measures did not possess a strict directional bias and were therefore explicitly a function of d and only implicitly a function of a.

A selected subset of these features were then used to classify 298 manually extracted posteriori interrib (intercostal) spaces on 95 zonal radiographs. The z = 50 percent jacknife testing results are shown in Table VI.

Table VI. Intercostal Image Test Results

		N	A
True Class	N	90	10
	A	5.2	94.8

Percent Correct 92.9

Concurrently work by Chien[16] also used the gray-level dependence matrix and kernels quite similar to those of Eqs. (31-35) as the basis for detection of venous hypertension. This disease is characterized by increased vascular profusion in the hilium with possibly decreased peripheral vascularity. The specific measurements used were of the form:

$$\bar{M}_k = \sum_i \sum_j A_k(i,j)p(i,j) \qquad k = 1,\ldots,5 \tag{39}$$

$$V_k = \sum_i \sum_j [A_k(i,j) - \bar{M}_k]^2 \; p(i,j)$$

In this case p(i,j) was a combined matrix of relative frequencies of distance d = 1 for all angles a. The $A_k(i,j)$ represented kernels almost identical to those of Eqs. (31-35).

A final set of features used in classifications were ratios of the following form:

$$R_k = \frac{\bar{M}_k^{\,p}}{\bar{M}_k^{\,h}} \tag{40a} \qquad \bar{R}_k = \frac{V_k^{\,p}}{V_k^{\,h}} \tag{40b}$$

$$D_k = \frac{\bar{M}_k^{\,p} - \bar{M}_k^{\,h}}{\bar{M}_k^{\,p} + \bar{M}_k^{\,h}} \tag{41a} \qquad \bar{D}_k = \frac{V_k^{\,p} - V_k^{\,h}}{V_k^{\,p} + V_k^{\,h}} \tag{41b}$$

As the lung was automatically divided into proximal and distal regions as shown in Fig. 11, the superscripts h or p refer to measurements made over the hilar or peripheral regions. A one-at-a-time jacknife test using a single edge

measure $A_k(i,j)$ with corresponding R_k or D_k, both classified 20 chest radiographs with a 90 percent accuracy.

Hall[35] used histogram and central-spatial moments to classify chest radiographs for presence of pneumoconiosis. The spatial moments were of the form

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q \, f(x,y) \tag{42}$$

where $\bar{x}$, $\bar{y}$ are centroids, and $f(x,y)$ represents lung zonal images. Spatial movement measurements were limited to a maximum of $p = 0,1,2$ and $q = 0,1,2$ or second-order type. In addition histogram $h(i)$ movements of mean (μ) standard deviation (σ) and coefficients of schemes γ_1, and excess γ_2 defined

$$\gamma_1 = \mu_3/\mu_2^{3/2} \qquad \gamma_2 = \mu_4/\mu_2^{2} \tag{43}$$

where

$$\mu_r = \frac{1}{N} \sum_i (i - \mu)^r h(i)$$

were computed for each zone. These features were computed for the entire lung zone, the horizontal gradient of the zone, and the intercostal region of the zone. Testing classification using a linear Gaussian classifier based on a jacknife of 50 percent yielded a 97 percent correct result. There were 38 total chest radiographs in the population.

A ten-dimensional subset of the lung texture measurements of Eqs. (31-38) and (42,43) has been used to classify the same 465-film test set and 130-film training set with the same lung-zone to film-diagnostic integration described in Eqs. (27,28).[36] The testing results are shown in Table VII.

Table VII. Digital Classification Results

		N	A	
Consensus Referred Truth	N	48 (49)	52 (53)	Percent Correct 74.0
	A	16 (58)	84 (305)	

The threshold was set at T = .56 as in the optical case. The results are not significantly different. However, the computer time necessary to compute the latter measures was longer.

A British group headed by Paton[37] has developed a single feature called the diversity or textural roughness measure for use in computer grading of

pneumoconiosis from the chest radiograph. Each intercostal space in the chest radiograph was divided into 300-square grids. Each grid was divided into nine squares. For each square the maximum of the rate of increase in density (the gradient) was calculated as one of eight possible directions. The diversity in a single grid was defined as the maximum number of directions taken by this gradient. The total diversity over all grids in the intercostal spaces is then a histogram. It was found that these diversity histograms broadened indicating an increasing diversity or textural roughness with increased severity of pneumoconiosis. A one-at-a-time jacknife using a linear classifier missed one film out of thirty-five. Correlation between the computed score and a consensus of 11 radiologists was .88. The correlation among individual radiologists and their consensus averaged .96.

Large Opacity and Vascularity Detection

The image-textural based approaches for lung disease detection discussed to this point have one element in common. All were designed to find characteristic measures of increased or decreased lung vascularity, or numerous small opacities which obscured this vascularity. The two examples to follow differ in that both attempted to detect and measure opacities and/or vascularity directly. These examples will use syntactual or structural pattern recognition as well as the previously described statistical pattern recognition approaches. Why has the structural approach been less explored? Most probably because it presents a more difficult problem. Vascular structure in the lung as projected onto the chest radiograph is more or less tree-like in nature with twin trunks emanating from the pulmonary arteries and veins. However, even in normal anatomy this structure is ill-defined, and smaller vessels are embedded in the fine irregular patterns that are often hardly visible. Toriwaki[38] views the nature of this problem as partially deterministic and partially stochastic. His detection of large tuberculosis opacities was predicated on the prior detection of much of the normal lung vascularity. This detection process consists of a top-down procedure of the sequential operations of linear filtering to extract lines of proper orientation, gray-level thresholding of the filtered image, followed by extraction of candidate vessel regions through the use of preset length and/or area thresholds. The major vessels near the hilium are detected with a deterministic decision rule based on a unanimous excessing of four preset thresholds. The smaller vessels are added to the larger ones using a statistical nearest-neighbor classification technique. Large suspect regions not intersected by detected vascularity are considered true large opacities. Figure 17 shows a right lung with traced vascularity and detected large opacities. Using 18 radiographs, 98 of 109 total suspect regions have been correctly classified as vascularity or opacity.

Ballard[15] has used a top-down ladder-structured decision tree approach shown in Fig. 18 to isolate coin lesions in the chest radiograph. The elements of the process involved the sequential operations of two-dimensional gradient analysis, followed by circularity detection to extract both suspect small-tumor regions and large-tumors per se. Next, dynamic programming was utilized to extract, if possible, a closed boundary of maximum edge sharpness and curvature. Surviving regions were then classified using a statistical nearest-neighbor decision rule utilizing extracted density, area, and boundary features within those surviving regions. This paradigm was tested on six radiographs. Three large tumors (greater than 2-cm diameter) were detected without error. Five more tumors 1 cm in size were also successfully detected.

Conclusions

This report has attempted to survey the field of computer detection and classification of chest radiographs. If one is to draw any broad conclusions,

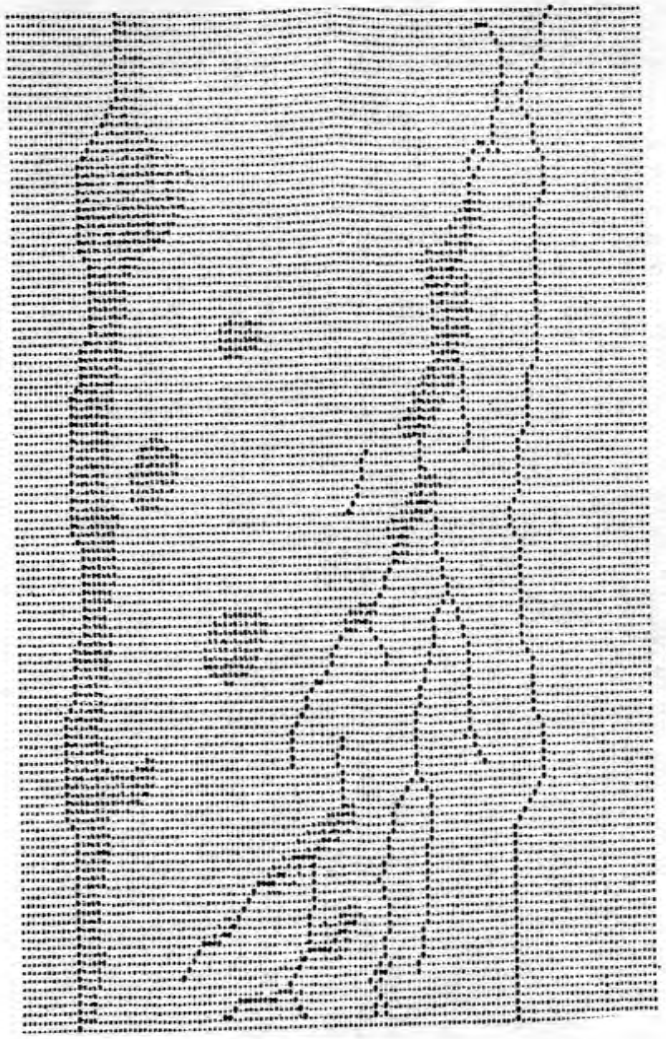


Fig. 17. Right lung with traced vascularity.

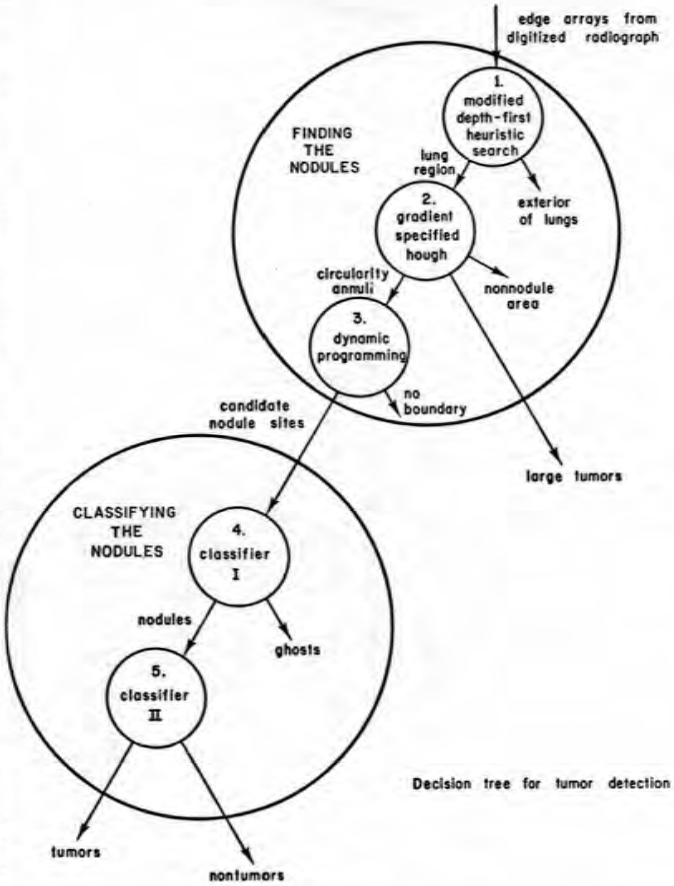


Fig. 18a. Ladder decision tree for tumor detection.

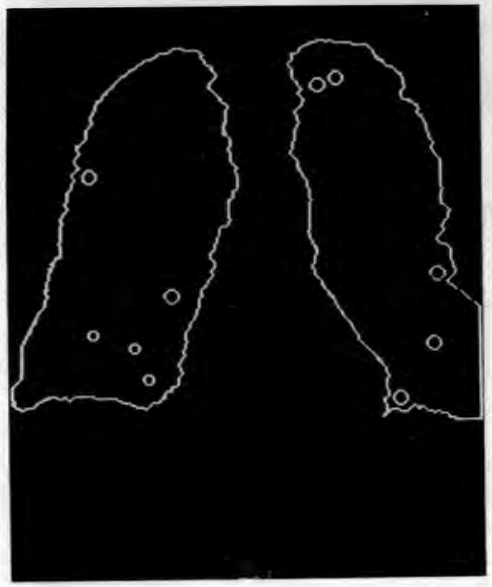


Fig. 18b. Detected coin lesions.

they would be that automated classification of chest radiographs for a general normal-abnormal diagnosis will be well in the future. However, the automatic detection and measurement of anatomic macrostructure such as lung, heart, diaphragm, and (for the most part) ribs, can be considered essentially complete. It is also highly probable that the heart and major mediastinal vessels can at present be reliably computer-classified into a normal or abnormal state.

Lung diagnosis is still an area where much research is needed. Many of the results presented in this report are predicated on a very small population of films and can best be described as feasibility studies. However, their instructional value should not be diminished by this fact. In addition, several of the textural features discussed would present serious diagnostic system throughput problems if implemented on the comparatively modest computational capability which would most probably be available for the purpose at present. For this reason, the inherent parallelism of optical texture-feature extraction techniques is attractive to the author over the short term when these are appropriate to the diagnostic task. Over the far term of 10 years or more, digital technology will undoubtedly shift the emphasis toward completely digital as opposed to hybrid diagnostic systems.

It would appear that diagnostic screening devices for a specific diagnostic purpose stand the highest probability of being successful over the next decade. Systems which hold high promise are the tuberculosis detection scheme of Toriwaki and the pneumoconiosis screening routines.

Acknowledgments

The author wishes to thank the many authors and several publishers who allowed their figures to be reproduced in this survey. Their cooperation is greatly appreciated.

REFERENCES

1. R. Brown et al., "Preliminary Estimates from U.S. Public Health Service, 1970 X-ray Exposure Study," Sept. 1971.

2. H. C. Becker et al., "Digital Computer Determination of a Medical Diagnostic Index Directly from Chest X-ray Images," IEEE Trans. Bio. Eng., Vol. 11, pp 67-72, July 1964.

3. R. P. Kruger, J. R. Townes, D. L. Hall, S. J. Dwyer III, and G. S. Lodwick, "Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors," IEEE Trans. on Bio. Eng., Vol. 19, pp 175-186, May 1972.

4. F. X. Roellinger, A. E. Kahvice, J. K. Chang, C. A. Harlow, S. J. Dwyer III, G. S. Lodwick, "Computer Analysis of Chest Radiographs," Tech. Report IAL-TR-20-73, Univ. of Missouri, July 1973.

5. N. Sezaki and K. Ukena, "Automated Computation of Cardiothoracic Ratio with Application to Mass Screening," IEEE Trans. on Bio. Eng., Vol. 20, No. 4, pp 248-253, July 1973.

6. E. L. Hall, R. P. Kruger, A. F. Turner, "An Optical-Digital System for Automatic Processing of Chest X-rays," Optical Eng., Vol. 13, No. 3, pp 250-257, May 1974.

7. E. L. Hall, R. P. Kruger, A. F. Turner, "Automated Measurements from Chest X-rays for Lung Disease Classification," Proc. of the Conf. on Computer Graphics, Pattern Recognition and Data Structures, pp 75-83, May 1975.

8. W. Spiesberger, "Heart Lung Border Outline in Chest X-rays," Proc. Second Joint Int. Conf. Pattern Recognition, Copenhagen, August 1974.

9. P. P. Tsiang, C. A. Harlow, G. S. Lodwick, "The Computer Analysis of Chest Radiographs," Tech. Report IAL-TR-30-74, April 1974.

10. E. L. Hall, "Almost Uniform Distributions for Image Enhancement," IEEE Trans. on Computers, Feb. 1974.

11. J. L. Paul, M. D. Levine, "The Measurement of Total Lung Capacity Based on a Computer Analysis of Anterior and Lateral Radiographic Chest Images," IEEE Trans. on Bio. Eng., Vol. 21, No. 6, pp 444-451, November 1974.

12. D. A. Ausherman, S. J. Dwyer, G. S. Lodwick, "Extraction of Connected Edges from Knee Radiographs," IEEE Trans. on Computers, Vol. 21, pp 753-758, July 1972.

13. H. J. Barnhard, J. A. Pierce, J. W. Joyce, J. H. Bates, "Roentgenographic determination of total lung capacity," Amer. Journ. of Medicine, Vol. 28, pp 51-60, January 1960.

14. D. H. Ballard, "Hierarchic Recognition of Tumors in Chest Radiographs," Tech. Rep. TR 74-4, Univ. of Calif. Irvine, June 1974.

15. D. H. Ballard, J. Sklansky, "A Ladder Structured Decision Tree for Recognizing Tumors in Chest Radiographs," Tech. Rep. TR-74-5, Univ. of Calif. Irvine, November 1974.

16. Y. P. Chien, K. S. Fu, "Recognition of X-ray Picture Patterns," IEEE Trans. on Systems, Man and Cybernetics, Vol. 4, No. 5, pp 145-156, March 1974.

17. Y. P. Chien, K. S. Fu, "A Decision Function Method for Boundary Detection," Computer Graphics and Image Processing, Vol. 3, pp 125-140, 1974.

18. J. Toriwaki, Y. Swenaga, T. Negore, T. Fukumara, "Pattern Recognition of Chest X-ray Images," Computer Graphics and Image Processing, Vol. 2, pp 252-271, 1973.

19. J. Toriwaki, H. Koshimizu, T. Fukumura, "The Software System for Automated Interpretation of Chest Photofluorograms," Digital Processing of Biomedical Images, Edited by K. Prestron, Jr. and M. Onoe, Plenum Pub. Corp., 1976.

20. H. Wechsler, J. Sklansky, "Finding the Rib Cage in Chest Radiographs," Tech. Rep. TP-75-8, Univ. of Calif. Irvine, November 1975.

21. H. Wechsler, T. Akers, K. S. Fu, "Image Processing Algorithms Applied to Rib Boundary Extraction in Chest Radiographs," Purdue Univ., Tech. Rep. 1977.

22. E. Persoon, "A New Edge Detection Algorithm and Its Application in Picture Processing," Computer Graphics and Image Processing, Vol. 5, pp 425-446, 1976.

23. L. H. Garland, "Studies on the Accuracy of Diagnostic Procedures," Amer. J. of Roent. and Rad. Therapy, Vol. 82, No. 1, pp 25-37, July 1959.

24. J. Jerushalmy, "The Statistical Assessment of the Variability in Observer Perception and Description of Roentgenographic Pulmonary Shadows," Radiologic Clinics of North America, Vol. VII, No. 3, pp 381-391, December 1969.

25. G. Jacobson, H. Bohlig, R. Keviluoto, "Essentials of Chest Radiography," Radiology, Vol. 95, May 1970.

26. R. P. Kruger, E. L. Hall, A. F. Turner, "A Hybrid Optical-Digital Radiography Based System for Lung Disease Detection," Applied Optics, October 1977.

27. H. C. Becker, P. H. Meyers, C. M. Nice, "Laser Light Diffraction, Spatial Filtering and Reconstruction of Medical Radiographic Images: A Pilot Study," IEEE Trans. on Bio. Eng., Vol. 15, No. 3, July 1968.

28. G. Revesz, H. Kundel, "Feasibility of Classifying Disseminated Pulmonary Diseases Based on Their Fourier Spectra," Investigative Radiology, Vol. 8, pp 345-349, September 1973.

29. C. Biagini, V. Cavallo, F. de Rosis, F. Mazzura, "Progress on Automatic Analysis of Chest Radiographs," Proc. XIII Int. Congress of Radiology, Madrid, October 1973.

30. R. P. Kruger, W. B. Thompson, A. F. Turner, "Computer Diagnosis of Pneumoconiosis," IEEE Trans. on System, Man and Cybernetics, Vol. 4, No. 1, pp 40-49, January 1974.

31. H. Stark, D. Lee, "An Optical Approach to Pattern Recognition of Coal Workers' Pneumoconiosis," IEEE Trans. on Systems, Man, and Cybernetics, Vol. 6, No. 11, November 1976.

32. D. J. Goodenough, "Need for Physical and Psychophysical Measures of System Performance," Proc. Med. X-ray Photo-Optical Systems Evaluation Symp., Columbia, MD, October 1974.

33. T. C. Strand, A. F. Turner, R. P. Kruger, H. I. Meyers, "Automatic Detection of Alterations in Pulmonary Venous Pressure," Proc. of the San Diego Bio. Symp., February 1977.

34. R. N. Sutton, E. L. Hall, "Texture Measures for Automatic Classification of Pulmonary Disease," IEEE Trans. on Computers, pp 667-676, July 1972.

35. E. L. Hall, W. O. Crawford, F. E. Roberts, "Computer Classification of Pneumoconiosis from Radiographs of Coal Workers," IEEE Trans. on Bio. Eng., Vol. 22, No. 6, pp 518-527, November 1975.

36. E. L. Hall, B. K. Rouge, R. P. Kruger, "Automated Chest X-ray Analysis," SPIE, Vol. 89, Applications of Optics in Medicine and Biology, pp 109-118, 1976.

37. J. R. Jagoe, K. A. Paton, "Measurement of Pneumoconiosis by Computer," IEEE Trans. on Comp., Vol. 25, No. 1, pp 95-97, January 1976.

38. J. Toriwaki, J. Hasegawa, T. Fukumura, "Recognition of Vessel Shadows for Automated Measurements and Classification of Chest Photofluorograms," Proc. Third Int. Conf. on Pattern Recognition Symposium on Computer-Aided Diagnosis of Medical Images, pp 1-8, November 1976.

PROCESSING THE DIGITISED RADIOGRAPHIC IMAGE:

CAN DECISION AIDING RANGES BE USED AS AN ALTERNATIVE TO ROC CURVES

N.J.G. Brown, A.T. Elliott and K.E. Britton
Departments of Nuclear Medicine, The Middlesex Hospital Medical School
and St Bartholomew's Hospital, London, U.K.

ABSTRACT

The ROC method involves the concept of the 'right' answer whereas in reality a spectrum of results is obtained in any set of patient data with no clear dividing lines. An alternative approach is to obtain quantitative indices from clinical data and to use decision aiding ranges to determine the course of action to be taken in a particular case. This method has been applied in the radionuclide brain scan. Its application to the diagnostic radiograph has been investigated by means of a rotating drum type digitizer interfaced to a nuclear medicine computer system. Illustrations are given of digitized radiographs taken of the skull, the pituitary fossa and the middle phalanx of the finger.

INTRODUCTION

The key problem in diagnostic imaging in a clinical context is the objective assessment of the worth of the result. The traditional method of patient management is to first obtain an accurate diagnosis of the disease and then proceed to treat this appropriately. The implied assumptions of this method are firstly that individuals can be unequivocally divided into sets of healthy and diseased individuals and secondly that tests are available which can be used to determine to which set a given individual belongs. Three factors give rise to problems when the attempt is made to allocate an individual to one set or the other: firstly, since individuals are not physiologically identical in all respects, there is often no clear division between the healthy and the sick; secondly, diagnostic tests have systematic and random errors associated with them and thirdly, different tests reflect different aspects of the individual - his anatomy, physiology or mental state, for example. (Furthermore, even within a group different tests reflect different aspects. For example, different physiological tests can demonstrate different aspects of organ function). Because of these three factors the results of different tests are often in conflict. When this happens the terms false positive and flase negative are used to describe some of the results no matter which of the three factors or combination of factors was responsible for the conflict. This conforms with the first assumption of the method that either unequivocally normal or abnormal states exist in all individuals. In other words there is always a 'right' answer in any clinical situation. Receiver operating characteristics, ROC, have been shown to be an

effective way of evaluating the effects of the second factor when the performance of different observers is compared.[1,2] Since it is required to know the 'right' answer, for each result studied, its practical application poses a major problem. This has limited its immediate general acceptance. Numerical techniques can be used to reduce systematic and random errors, and have been applied mainly to simple phantoms[3] and phantoms set in a 'clinical' background.[4] ROC curves are then created to assess observer variation. An alternative approach[5,6,7,8] is to attempt to translate the data into a format whereby strategies, such as those developed for the interpretation of the results of in vitro tests,[9] can be used.

In this method, the concept of the unequivocally defineable diagnosis of diseased states is abandoned in favour of the concept of differential management. The task of the clinician is to decide between the various possible management decisions in a given clinical situation. This does not require the specific definition of diseased states, but rather a set of operational diagnoses. This concept of decision aiding ranges[8,9] is based on three principles.

Firstly, relevant diagnostic tests are those whose results alter clinical management. Therefore the use of the results of such a test, whether in vitro or in vivo, should only be such as to decide between the different possible managements - hence decision aiding ranges. Secondly, the acceptable error of any measurement can be defined as that which is required to distinguish between alternative decision ranges. For example, in the measurement of individual renal function in surgical practice, it is required to select between a range of less than 10% of total function leading to a decision to undertake nephrectomy and over 25% of total function leading to a restorative operation. Between 10% and 25% of total function, the decision to resect or restore is modified by other factors such as the presence of pain, infection, complex anatomy etc. Thirdly, decision aiding ranges can be used to divide results into five categories, S1-S5 as a way of redefining 'normal' and 'abnormal'. S1 and S5 define definitely abnormally low and high ranges of values into which no normal person should ever fall. S3 defines a range of values into which no definitely abnormal person should ever fall. S2 and S4 define two clear categories of 'borderline' results into which a normal or abnormal patient might fall. The decision then is to apply a second more specific test for each borderline category to distinguish further between the normal and the abnormal (e.g. estimation of thyroid stimulating hormone, TSH for the borderline low (S2) result of a Free Thyroxine Index[9]).

We[5,6,7,8] have shown that by 'quantitation of difference' using an interactive computer system, S3, S4 and S5 categories can be defined numerically for any part of any view of a brain scan. In this technique a quantitative estimate is obtained of the difference between the observed and expected count rates at a suspect point. The most important application of the brain scan is to demonstrate the absence of abnormalities in the brain. In our experience only 5% of patients referred from general physicians and surgeons have S5 (definitely positive) brain scans compared with 50% of patients referred from neurosurgeons. These results indicate that S3 is the most important decision aiding range for brain scanning. Indeed, following Bayes theorem, these statistics imply that, as CAT scanning is a better test for demonstrating pathology, it would be more appropriately applied only to neurosurgical referrals with their high prevalence of positive results. It also would be appropriate for S4 brain scans, which represented 15% of general referrals. Normal (S3) results occurred in 80%. This data was confirmed by a retrospective assessment of 101 histologically proven pathological brain tumors whose previous scans when analysed by quantitation of difference showed 84% in the S5 category and the other 16% in the S4 category.

Houston et al[4] recently demonstrated that the quantitation of difference technique was the most successful of three methods of brain scan quantitation.

In summary decision aiding ranges avoid the difficulties with the conventions: normal, abnormal, false positive, false negative, inherent in the use of ROC curve theory when this is applied to assess the value of a given test when used in a particular way.

Given the concept of decision aiding ranges, it was felt appropriate to attempt to apply it to the digitised radiographic image to see if it could aid the discrimination between the normal and abnormal pituitary fossa and the normal and abnormal middle phalangeal bone of the index finger in the objective assessment of the diagnosis and progress of hyper parathyroidism in response to treatment. The remainder of this paper concerns work showing the technical feasability of such a project. During this work we have demonstrated the potential value of the use of enhancement procedures in diagnostic Radiology.

METHODOLOGY

The computed radioisotope image is obtained from a digital record of the convolution of the imaging device measuring characteristic and the gamma ray flux generated by radioactive material distributed in the body, modulated by statistical noise. In the case of the radiograph the record is of shadows cast by identifiable physical structures with negligible statistical error modified by the effects of scatter. In interpreting a radiograph it is required to infer the presence or absence of a physical feature on the basis of visual examination of the information in the recorded image. As has been previously argued[6] the eye is bad at estimating absolute differences in density. Furthermore there is a distinction between a primary physical defect, with an anatomical and pathological basis and an image defect which may or may not reflect the physical defect. If quantitative measures of image features can be made then, by performing studies in a clinical trial, the probability of a physical defect being present in the presence of a given quantitated image defect can be determined.

A Joyce Loebel Scandig 2 high speed rotating drum microdensitometer has been used to study radiographs. The specimen to be scanned, of maximum area 225 x 225 mm, is attached to a dynamically balanced drum. A small modification to the clamps will allow it to accommodate a 10" x 8" radiograph. A tungsten halogen light source with a 200 micron primary slit illuminates the film via a Kohler optical system to ensure uniform illumination over the digitized area. Light transmitted through the specimen is imaged by a fixed microscope objective onto a solid-state photodetector through a selectable measuring aperture of 25, 50, 100 or 200 microns. The optical density range is 0 to 3, which is output to the computer in 8-bit binary form, giving a density resolution of 1 part in 256: the density repeatability is 2 in 256, with a linearity of 1 in 256 from the best straight line fit. Once per revolution, the measuring system is automatically zero corrected when light is transmitted through a slot in the rotating drum.

The area to be scanned is selected on the Scandig through four sets of thumbwheel switches, as is the scan increment. Circumferential position (Y) is measured by counting the pulses from an incremental shaft encoder after its zero reference pulse, while axial position (X) is measured by counting the step rotations of the precision leadscrew which drives the optical system carriage, giving high positional accuracy.

In addition to the density data, scan increment (Y), end-of-line, data ready, stepping complete, end-scan and end-stop signals which are output to the computer; step (X), step direction (positive or negative), scan, stop and back-to-datum commands may be input from the computer.

Ordinarily, both the measuring aperture and scan increment are set to 200 microns since this has been found to be fine enough for the radiographs studied so far. Under software control, image element spacings of 400 and 1000 microns can also be obtained by ignoring the required data points from the microdensitometer. Using a 1000 micron increment, a specimen area of 25 x 25 cm can be digitised into a 251 x 251 matrix which is displayed for the observer to locate any regions of interest. These may then be rescanned at higher resolution.

To digitise part of a film the area limit switches on the Scandig unit are set to the maximum area and selection of the region to be digitised is accomplished by inputting from the VDU the values of x start, x stop, y start and y stop in mm. Under software control, data from outside the selected region is ignored. When scanning, the data rate is 20 kHz and a 225 x 225 mm area can be scanned in 100 seconds at a resolution of 200 microns.

Over the greater part of the density range of the film the density is proportional to the log to the base ten of the exposure and

$$D = K_1 + \lambda \log_{10} E \quad , \tag{1}$$

where D is the density K_1, λ are constants, and E is the exposure. Since the exposure depends on the number of photons arriving at the film we may assume:

$$E = K_2 N_0 \, e^{-\mu d} \quad , \tag{2}$$

where E is the exposure produced when n_0 X-ray photons are incident on a slab of material between source and film of thickness d and mean absorption coefficient μ, K_2 being a constant.

Substituting in 1 yields

$$\begin{aligned} D &= K_1 + \lambda \log_{10}(K_2 N_0 e^{-\mu d}) \\ &= K_1 + \lambda \log_{10} K_2 + \lambda \log_{10} N_0 + \lambda \log_{10} e^{-\mu d} \\ &= K + \lambda \log_{10} e^{-\mu d} \quad , \end{aligned}$$

where $K = K_1 + \lambda \log_{10} K_2 + \lambda \log_{10} N_0$

$$= K + \frac{\lambda \log_e e^{-\mu d}}{\log_e 10} \quad .$$

Hence

$$(D - K) \frac{\log_e 10}{\lambda} = -\mu d \quad .$$

So

$$\mu d = B(K - D) \quad ,$$

where

$$B = \log_e 10 / \lambda \quad . \tag{3}$$

When μd is zero (no absorption) then D is a maximum value D_{max} therefore

$$K = D_{max}, \text{ so } \mu d \, \alpha (D_{max} - D)$$

When inputting the density values from the digitizer the computer subtracts each value from D_{max} and stores the result in the appropriate point of the image matrix. Thus the values stored are proportional to mean μd value for each point. Tests have been performed with an aluminium test step wedge which show that the characteristic is linear down to densities of around 0.6. Below this the rate of change of density with increasing d falls off and the curve is substantially flat below 0.25. The shape of the curve changes somewhat with the presence of scatter material which raises the background level. However an approximate correction for the non-linearity within the density range 0.6 to 0.25 can be made. This is done optionally using a lookup table by the software which reads the data in from the digitizer. (Fig. 3)

The digitizer is interfaced via CAMAC to the Department's old Varian 620i computer which is of 24K memory capacity and provides disc storage for image matrices of any size. Also interfaced via CAMAC is a Princeton PEP 400 scan converter grey scale TV display system. Three DACs are used: one for X deflection, one for Y deflection (each of 10 bit accuracy) and one for Z(brightness) of 7 bit accuracy. In practice the two least significant bits of the deflection DAC are not used. The X and Y deflection gains have been adjusted so that the top 8 bits yield a 256 x 256 matrix whose dots just touch to give a uniform grey level in a uniform area.

The system is controlled by a VDU and switch panel. When the grey scale TV program is called it first reads all the data from the selected disc file and finds the minimum and maximum values, termed zero and 100 percent values respectively. When a button on the control panel is pushed the computer reads minimum and maximum percentage values to be displayed from two sets of edge switches. An image is then written to the storage tube in which the brightness of each point (in the range 0 to 128) is obtained by linear interpolation between the data values corresponding to the minimum and maximum values required. Thus upper and lower threshold control is provided and the full range of the scan converter is always used.

The same edge switches are used to control threshold and percentage colour range control for the displays of images up to 128 x 96 on an Elscint VDP. The program is similar to that described elsewhere[5,6] in which the range of each colour is computed as a percentage of the threshold value allowing the difference between the threshold value and any point in the image to be readily found. It has been found convenient for radiographs to add an option whereby the colour ranges may be made to extend downwards from the threshold value instead of upwards.

Conventional image processing procedures such as smoothing and unsharp masking[11] may be readily applied to images of up to 80 x 80 elements on our statos plotter and Tektronix 611 displays. Software modifications are under development whereby the 1.2 Megawords of data obtained by sampling the entire 225 x 225 mm area of film scanned at 200 microns can be transferred to disc. An image of this will be displayed at 1 mm resolution and a bug used interactively to select a point of interest. The software will then select data required from the disc so as to produce a full size display centred on the bug at the selected resolution down to 200 microns. It is also proposed to replace the scan converter and existing colour displays by a 256 x 256 system with a built-in memory which will enable the effects of changes in contrast and threshold settings in both black and white and colour to be displayed virtually instantaneously, by changing values in the hardward lookup tables. The

existing system being unacceptably slow at the 256 x 256 matrix size. Each image of this size requires seconds to be written to the scan converter in spite of the use of a software lookup table technique for determining the brightness in which a given value is to be displayed.

With the system in its present form it is possible to bring in, correct for non linearity and display (with variable threshold and contrast) grey scale images of up to 256 by 256 elements in size. Images up to 128 x 96 may be displayed in colour and quantitative statements about the percentage difference in mean d values between different points made by the methods previously described[5,6]. In addition images of up to 128 x 128 elements may be subjected to traditional smoothing and enhancing procedures such as unsharp marking[11].

RESULTS

Results obtained with the system are shown in Figs. 1 - 9. The interactive colour display system was used to obtain quantitative data from the radiographs of Fig. 9.

DISCUSSION

Figures 1 - 6 demonstrate the effects of digitising and enhancing some clinically important parts of a skull radiograph. Quantitative methods have been studied in the treatment of hyperparathyroidism as shown by radiographs of the finger. Quantitative values were obtained from the data of figure 9 as follows; since the proximal phalanx is relatively unaffected the point of least absorption by the trabecular bone of this phalanx, about 1 cm clear of the joint space was taken as the reference point. By interactive use of the controls the threshold was set to the value at this point. Before treatment the abnormal bone was found to be 27% down in μd value from the reference value. However, after treatment it was only 18% down from the reference value, i.e. a 9% or one third improvement in the original 27% deficiency.

We have shown therefore, that quantitative sequential studies in the same patient can be usefully performed. Absolute measurements, however, will require that a step wedge be included on the radiograph so that the effects of different voltage and beam currents may be normalised.

These methods may in the future help to resolve the continuing conflict between what is a radiologically normal, borderline and abnormal pituitary fossa and aid the development of decision aiding ranges for this. The assessment of the presence or absence of hyperparathyroid changes in the phalanges and their response to treatment will, we hope, also prove useful.

Fig. 1. Original Skull X-ray

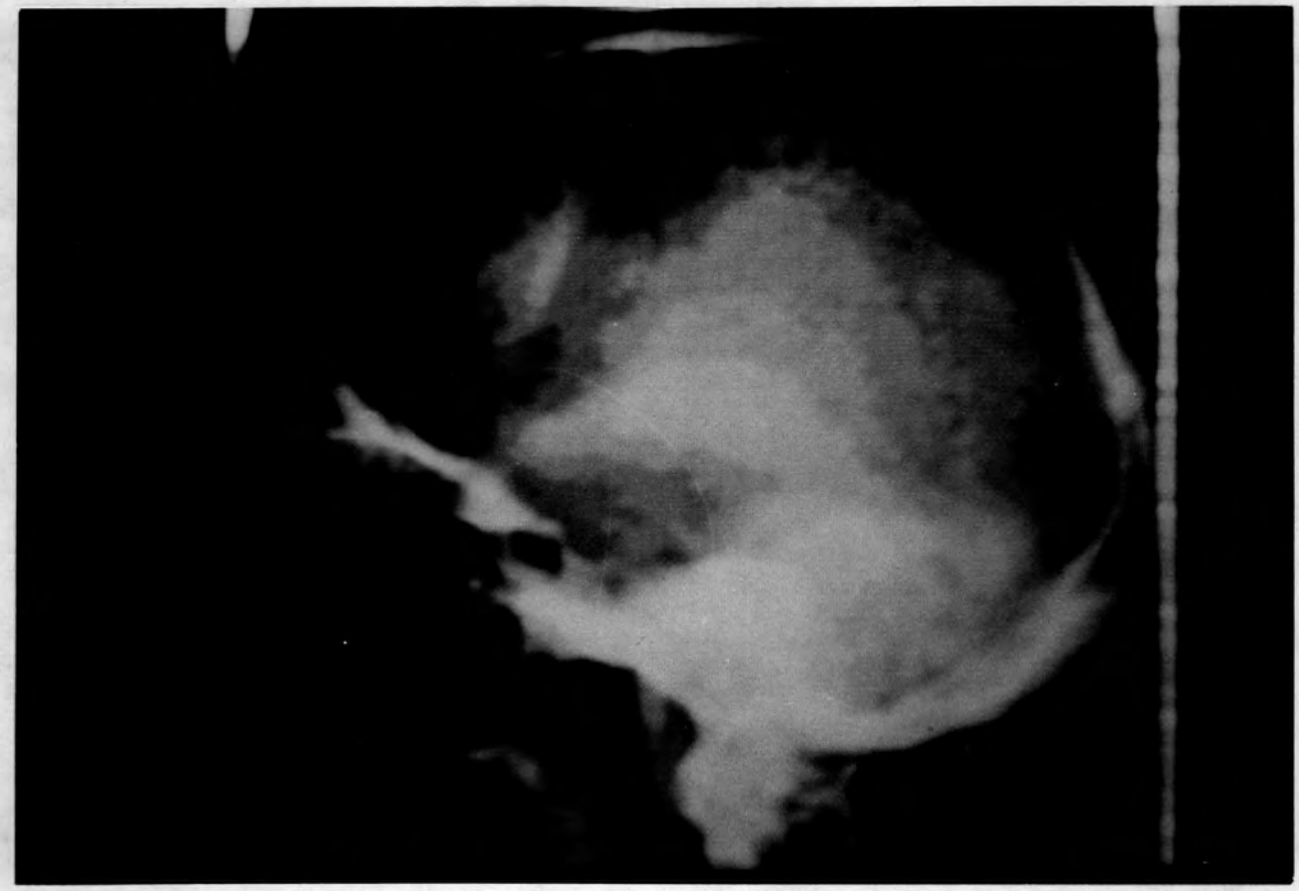


Fig. 2. Whole Skull, digitised at a resolution of 1000 microns

Fig. 2-a) displayed range 0 - 99%

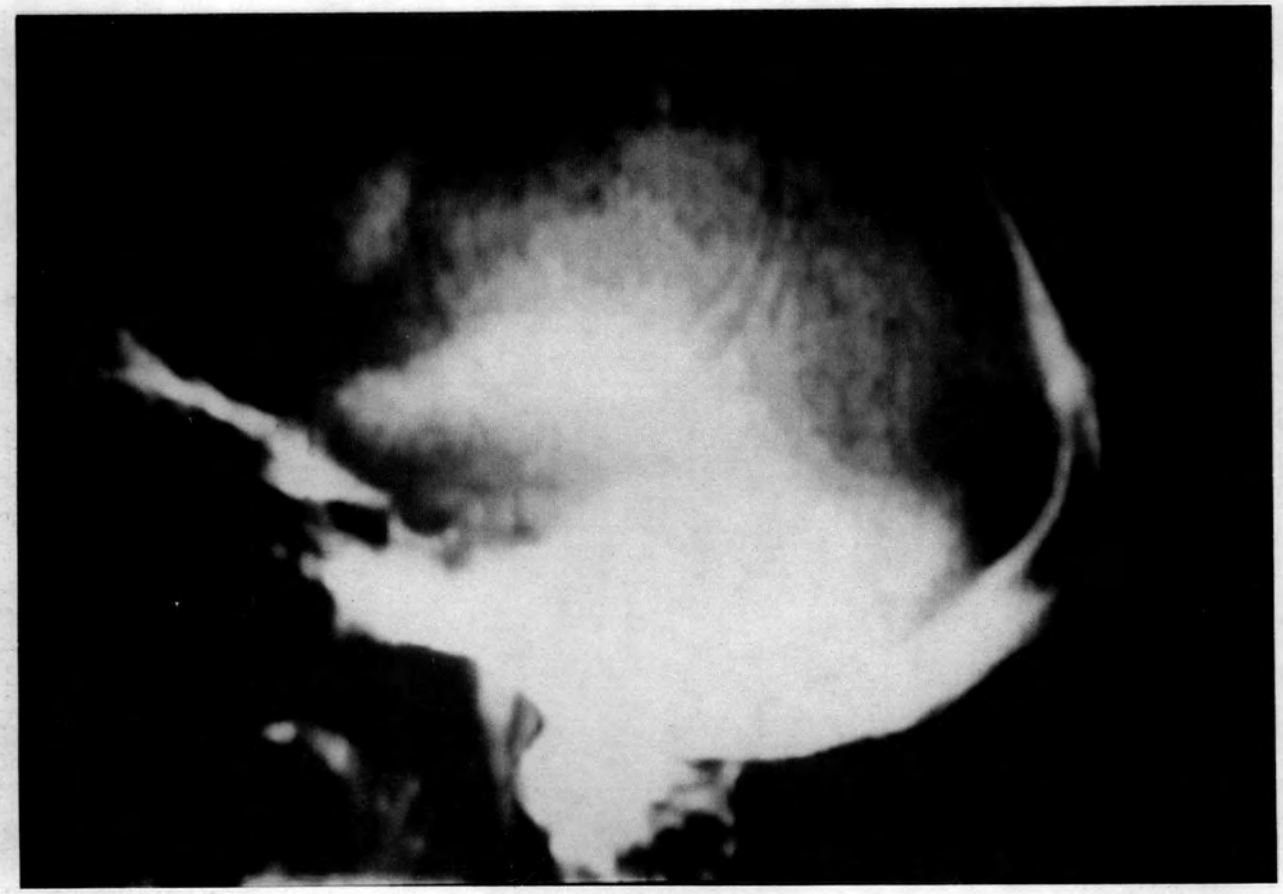


Fig. 2-b) displayed range 30 - 99%

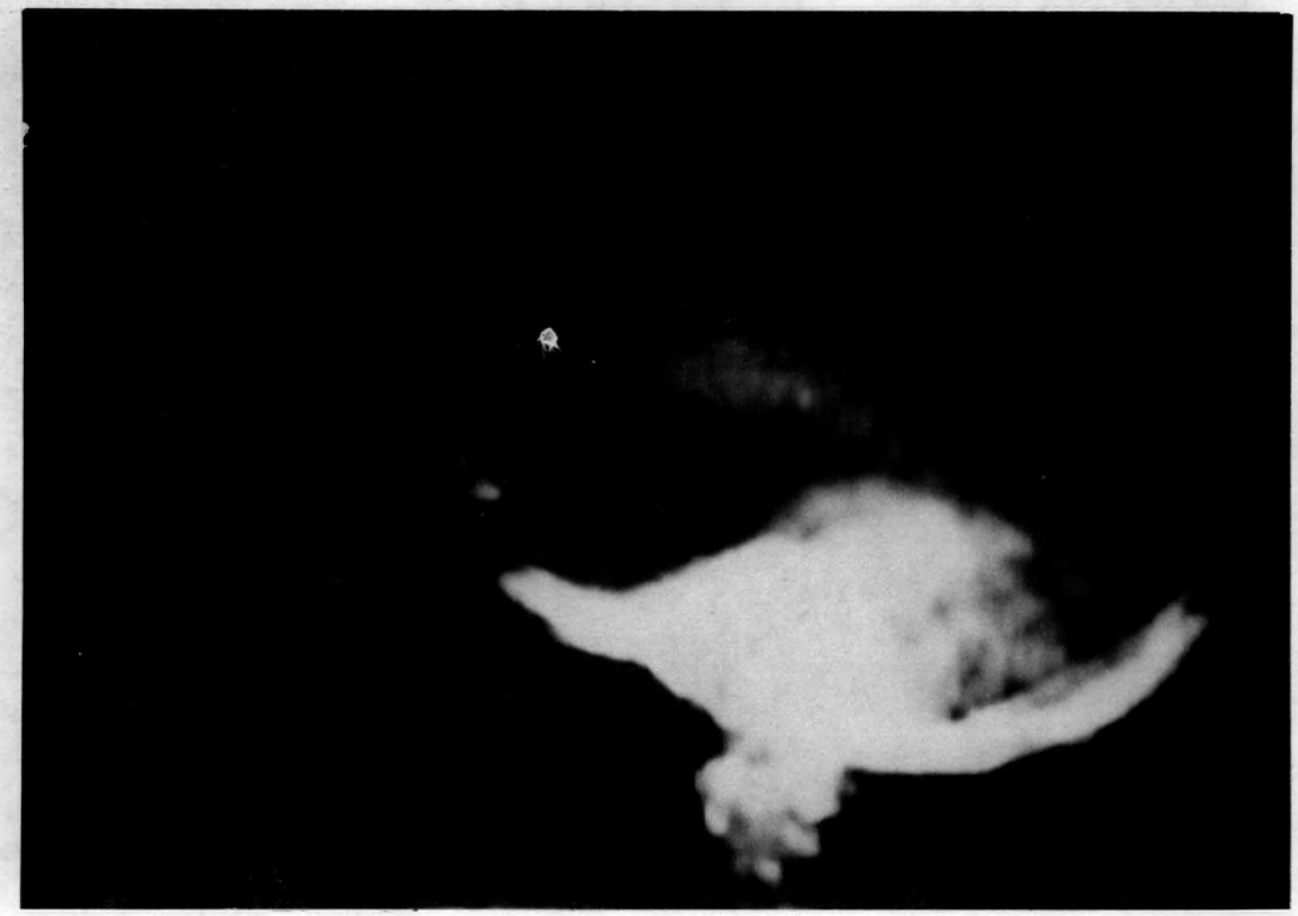


Fig. 2-c) displayed range 60 - 99%

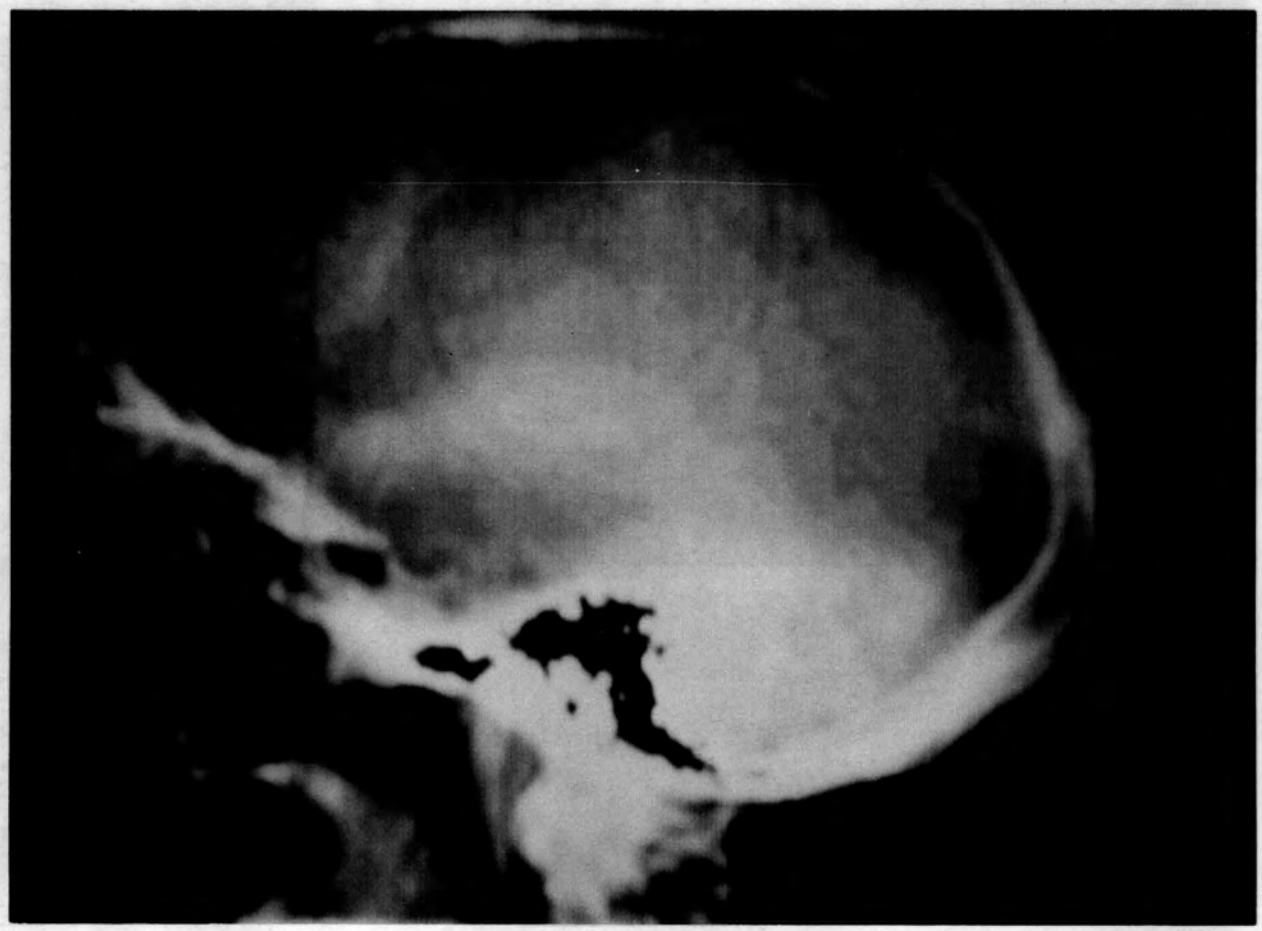


Fig. 2- d) displayed range 0 - 85%: the region in black, of μd values betweem 85 and 100% of the maximum, is the petrous temporal bone. Note also the pituitary fossa.

Fig. 3. Section of skull, encompassing the petrous temporal bone and pituitary fossa: resolution 1000 microns.

a) not enhanced (i.e., no linearity correction); displayed range 50 - 99%

b) enhanced (i.e., linearity correction), displayed range also 50 - 99%; note greater visual detail. (Black regions in bone are outside correctable range.)

Fig. 4. Section of skull, petrous temporal bone; digitised at a resolution of 400 microns.

a) not enhanced, displayed range 80 - 99%

b) enhanced, displayed range also 80 - 99%. Suggests semicircular canals at lower right

Fig. 5. Pituitary fossa original

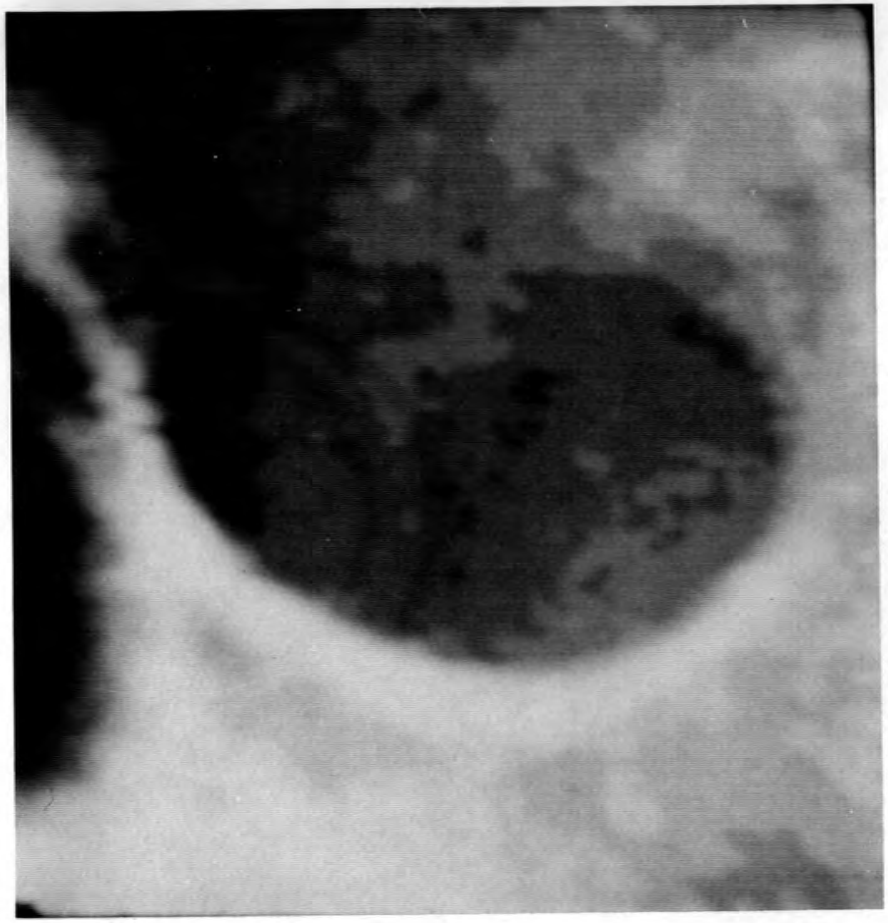


Fig. 6. Section of skull, pituitary fossa- digitised at a resolution of 200 microns.

a) not enhanced, showing suggestion of internal detail; displayed range 55 - 99%. (Dark rings are scan converter artifacts.)

b) after 'Canterbury' processing: smoothed 3 times, ($^{13}/_{16}$th subtraction factor. Note that the internal features are now defined, such as the lobes of the pituitary as is structure in related bone: displayed range also 55 - 98%.

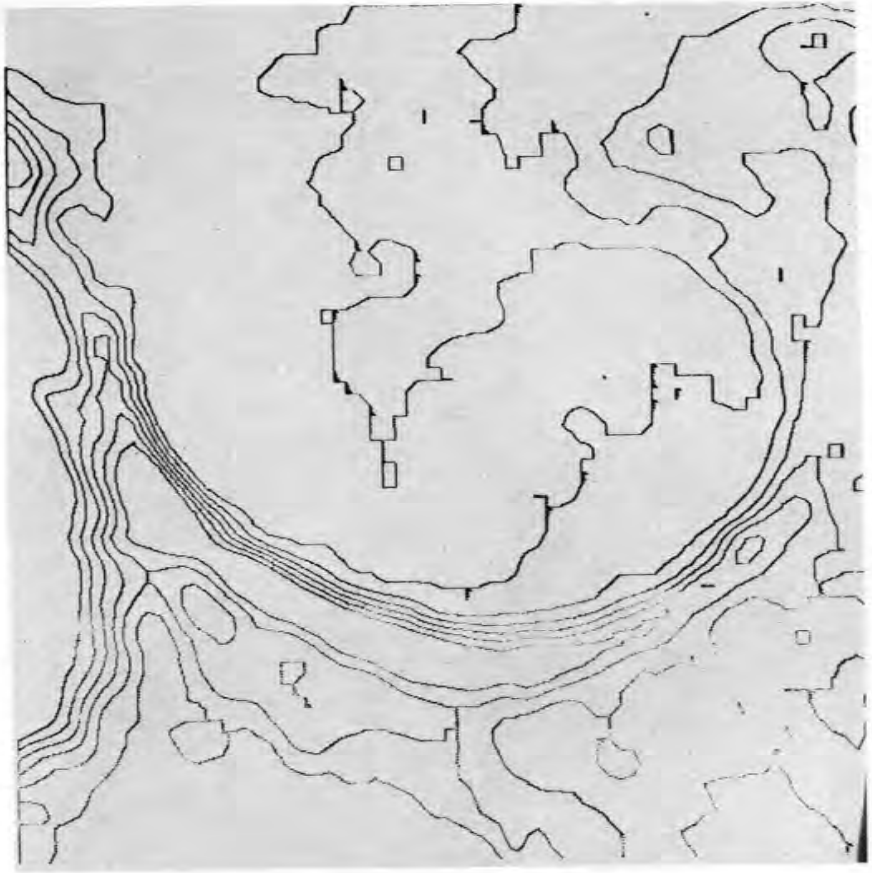


Fig. 6-c) data as 5a contour map

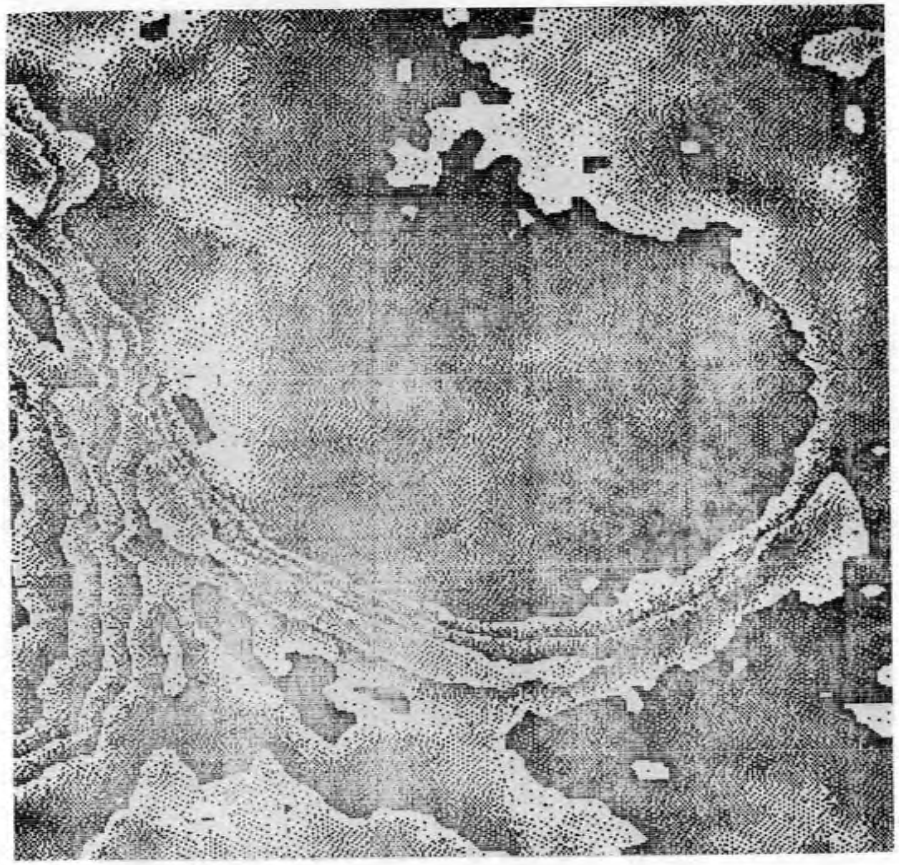


Fig. 6-d) data as 5a pseudo random dot grey scale

Fig. 7. Hand: original X-ray of 8(b)

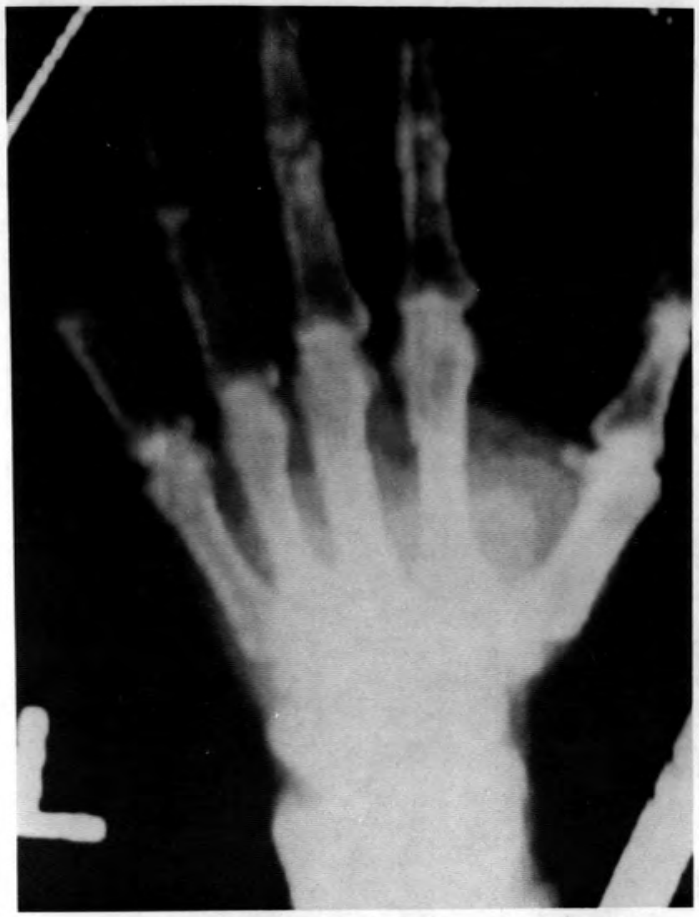


Fig. 8. Hand (of fig. 7) digitised at a resolution of 1000 microns - disease affects primarily the middle phalanx of the second and third digits, while the proximal phalanx is relatively unaffected.

a) displayed range 0 - 99% b) displayed range 20 - 75%

Fig. 9. Section of third digit, including joint area between proximal and middle phalanges; digitised at a resolution of 200 microns.

a) abnormal, prior to treatment: displayed range 70 - 89%; note lesion on radial edge of middle phalanx.

b) abnormal, after treatment: displayed range 50 - 89%.

REFERENCES

1. L.B. Lusted, Introduction to Medical Decision Making, Charles C. Thomas, Springfield, (1968).

2. C.E. Metz, S.J. Starr, L.B. Lusted and K. Rossmann, Progress in evaluation of human observer visual detection performance using the ROC curve approach Proceedings of the 4th International Conference on infarction processing in scintigraphy, A. Todd-Pokropek and C. Raynaud editors, Commisariat a l'energie atomique, Orsay, Paris, 1976.

3. P. Sharp and J. Mullard, The measurement of the performance of the display system of a radioisotope imaging device: the multi-element band display, Brit. J. Radiol., 49, 270, (1976)

4. A.S. Houston, An intercomparison of three quantitative data processing techniques using mathematical tumours and receiver operating characteristic analysis, In Proceedings of the International Symposium on Medical Radionuclide Imaging, Los Angeles, IAEA, Vienna in press.

5. N.J.G. Brown, T. Budd and K.E. Britton. On the transmission of quantitative difference information via an interactive computer system to the clinical interpreter, p. 377 in Proceedings of the 4th International Conference on information processing scintigraphy. A. Todd-Pokropek and C. Reynaud, Editors, Commisariat a l'energie atomique, Orsay, Paris, 1975.

6. N.J.G. Brown, K.E. Britton and F.R. Cruz. Realising the quantitative potential of the Radioisotope Image in Proceedings of the International Symposium of Medical Radionuclide Imaging, Los Angeles, IAEA, Vienna in Press.

7. K.E. Britton, F.B. Cruz and N.J.G. Brown, Clinical Use of quantitative brain imaging, Proceedings of the 13th International Congress of Internal Medicine, Helsinki, S. Karger, Basel, in press.

8. K.E. Britton, Development of effective diagnostic strategies in quality control in Nuclear Medicine, B.A. Rhodes, Editor, Mosby Co. in press.

9. K.E. Britton, V. Quinn, B.L. Brown and R.P. Ekins, A Strategy for Thyroid Function Tests, Brit. Med. Journal, 3, 350 (1975).

10. K.E. Britton, Renal radionuclide techniques in their clinical context. Proceedings of the International Symposium on Medical Radionuclide Imaging, Los Angeles, IAEA, Vienna in press.

11. J.R. Corfield, Development of a fast image enhancement filter for routine use. p. 72.1 in Proceedings of 13th Annual Meeting of the Society of Nuclear Medicine, Fadl.s Forlag Copenhagen, 1975.

DISCUSSION:

A spirited discussion focused on the issue of whether Decision-Aiding Ranges (DAR's) as defined by Brown were to be considered as an alternative to ROC analysis or whether they were, rather, useful in a complementary way. C. Metz suggested that DAR's and ROC curves are directed toward different purposes: ROC analysis is a method for the retrospective evaluation of decisions, while DAR's provide a tool for making decisions. Agreeing, D. Goodenough pointed out that practical techniques for generating ROC curves require the decision-maker to grade either numerical test data or subjective confidence estimates into categories, which are strictly analogous to DAR's. A. Todd-Pokropek and Metz pointed out that the choice of boundaries defining the DAR's will influence the usefulness of the DAR approach and that the problem of evaluating the DAR approach remains. Brown and K. Britton (co-author of the presentation) agreed and suggested that ROC analysis might be used for that purpose. In response to a question from W. L. Rogers, Britton stated that they had done a retrospective evaluation of the DAR method as applied to interpretation of thyroid diagnostic data.

D. Janney remarked that in the course of the presentation, Brown had shown a number of displays that appeared to a layman to enhance a considerable amount of image detail. He asked if the approach had yet gained acceptance by clinicians to the extent that they use these displays as an aid to patient management. Brown replied that in radionuclide imaging the answer was "yes" but that this acceptance was not yet evident in diagnostic radiology. Brown emphasized that the fundamental point of the approach was not image enhancement as such, but rather presentation of quantitative information regarding differences in image density.

NEW METHODS FOR PULSATILE BLOOD FLOW MEASUREMENT USING CINE-DENSITOMETRY+

U. Efron*, Dan Kedem**, Drora Kedem**, C.W. Smith,
J. Erickson, R.R. Price, J.H. Nelson, and A.B. Brill
Vanderbilt University Medical Center
Nashville, Tennessee U.S.A.

ABSTRACT

Experimental measurements of pulsatile blood flow were carried out using standard cine-angiography techniques with subsequent digitization and computer processing of the cine film. New approaches to measurement of blood flow were examined. The commonly used center of gravity method for the bolus transit time determination is shown to be subject to serious error when backflow is present. An analytic method for the determination of instantaneous blood velocity has been developed and tested experimentally using controlled constant and pulsatile flow systems. The results show good agreement with the known flow values. The method was applied to cine recorded data of arterial blood flow in dogs. The results are in good agreement with the known shape of the velocity distribution curve in the normal aorta.

+This research was supported in part by SCOR grant HL-14192.

*On leave from Soreq Nuclear Research Center, Yavne, Israel.
**Soreq Nuclear Research Center, Yavne, Israel.

INTRODUCTION

Previous cine and video densitometry methods of measuring blood flow have shown quite scattered results for pulsatile systems whereas the correlations in constant-flow systems were usually quite good. (1-4)

The method employed consisted of cine and/or video tape recording of the x-ray image of the rapidly injected intraarterial radiopaque contrast dye. The subsequent analysis of the recording was aimed at obtaining the transit time of the bolus between two sites or "windows" along the blood vessel. This was done by observing the frame to frame progress of a well-defined point in the bolus density curve such as the leading edge or the peak point. A commonly employed method uses the time center of gravity (C.O.G.), of the density curve obtained at the two sites. The transit time is then taken as the time difference between these two centers of gravity time.

This report presents an analysis of both the currently used (C.O.G.) method and a newly developed instantaneous velocity method for the determination of blood flow.

Blood flow simulation experiments were carried out under both stationary and pulsatile flow conditions. Animal studies are also reported. The mathematical basis for both data processing methods is presented (i.e. the "conventional" C.O.G. method and the newly developed "Instantaneous velocity" method). The experimental data are analyzed using both methods, and the results compared.

Good correlation with the known flow values is found using both methods under stationary flow condition. In pulsatile flow conditions, poor correlation is found for the C.O.G. method whereas the instantaneous velocity method produces results with good correlations. It is verified that the new method correctly determines the direction of flow in cases where a visually observed backflow is present. The velocity distribution data obtained by applying the instantaneous velocity to dogs' aortic flow data compares well with the expected aortic blood flow distribution in normal dogs.

Finally, a comparison of the two methods for the determination of the average flow is carried out. In particular, an analysis of the limitation in applying the C.O.G. method to pulsatile flow condition is presented.

METHODS

Experimental Setup

The experiments were conducted in the Cardiac Catheterization Laboratory. at Vanderbilt University Hospital. The experimental setup is shown in Figure 1.

The x-ray tube operated at 50kv was pulsed synchronously with the cine camera at 60 c.p.s. with pulse durations of 2-5 msec. One (1) ml. of contrast dye -- "Renografin - 76" -- was injected using a programmable injector, at an injection speed of 59cc/sec. Experiments included both simulated blood flow studies and studies in the descending aorta in dogs.

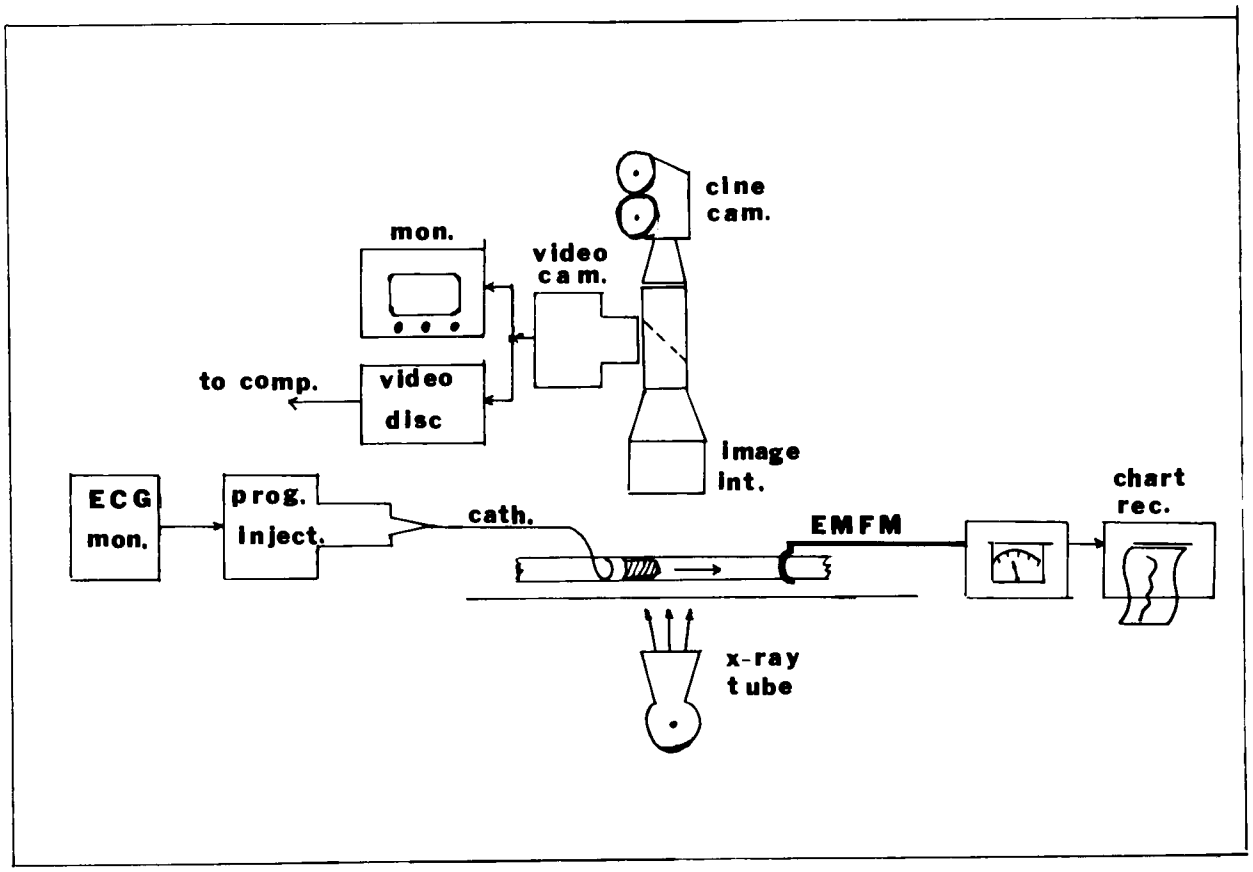


Figure 1. The experimental set up.

The experimental calibration studies used constant (gravitational) flow and pulsatile flow systems. In the constant flow simulation study, saline flowing under the influence of gravity controlled by a valve at the outlet of the container was used. The pump used in the pulsatile flow was a Travelol rotary pump with a cycle time of approximately 810 msec. Injections at several flow rates were made at the beginning and end of a pump rotation cycle. For flow corresponding to average velocities of 20 cm/s or less, backflow was visually observed. The flow in both continuous and pulsatile simulation experiments was measured using time sampled flow into a graduated cylinder. The tubing used was p.v.c. type with an inside diameter = 4.76 mm.

Dog studies were performed on anesthesized animals. A pigtail type catheter inserted into the abdominal aorta above the ostium of the right renal artery. Several injections were made throughout the cardiac cycle using the E.C.G. triggering capability of a programmable injector. The 35 mm cine film type was CFS Kodak. It was developed in an Oscar Fisher G-12 Developer.

Data Processing System

The cine film was mounted on a film transport device and trans-illuminated by diffused light. The image was digitized using a vidicon camera (Cohu 810) coupled to a video compressor (Colorado Video 270). The vidicon resolution was about 800 lines. The digitizer frequency was 10 MH_Z. Digitized data were transferred to a PDP-9 computer (Digital Equipment Corporation) using special routines to digitize selected regions of interests (usually lines across the vessel) and to store the total logarithmic intensity at the selected sites. The digitization of one frame (10 sites of ~ 100 point each), required about 10 sec. The number of frames digitized were 30-60 for a typical study.

The methods used for blood-flow determination

The method of the center of gravity (C.O.G.)

This method is based on the determination of transit time of the bolus between two sampling sites located along the vessel. The transit time is determined by calculating the time difference between the centers of gravity of the density curves at both sites.

The mathematical expression for the transit time obtained by using the C.O.G. method is given by:

$$\Delta t = t_B - t_A = \frac{1}{S_B}\int C_B(t).t.dt - \frac{1}{S_A}\int C_A(t).t.dt \ldots (1) \quad ,$$

where the "B" site is located downstream with respect to "A". $C_A(t)$, $C_B(t)$ are instantaneous densities of contrast dye observed respectively at sites "A" and "B".

S_A and S_B are the total (integral) amount of dye at "A" and "B" respectively.

Usually 6-10 equally spaced sites were digitized -- the program then proceeded to calculate the C.O.G. of the density curves at the sites analyzed using the data from the sequence of frames.

The "Instantaneous velocity" method

The density curves at multiple sites along the vessel possess spatial and temporal information related to the bolus density function $B(x,t)$. In order to make use of this information, we developed the technique which is called the instantaneous velocity method.

As the bolus is carried downstream, its profile is gradually changed due to intermixing with blood. However, for closely adjacent sites (i.e. several millimeters apart in an average blood velocity of 20-30 cm/s), one would expect the bolus shape to appear almost identical at both sites. The general mathematical expression for a physical phenomenon $B(x,t)$ moving at a constant speed V_0 in the +x direction, without changing its shape is:

$$B(x,t) \equiv B(x-V_0 t) \quad . \qquad (2)$$

(This expression states that a density observed at point 0 at time 0 will be observed at point x and time t where $x = V_0 t$.)

In the pulsatile case, the velocity is a time-dependent quantity, and hence the above expression must be replaced by:

$$B(x,t) \equiv B(x-I(t)) \quad , \qquad (3)$$

where

$$I(t) = \int^{t} V(t)dt \quad . \qquad (4)$$

The form (2.3) of the bolus function $B(x,t)$ leads to the following relation between the spatial and temporal derivates of $B(x,t)$:

$$\frac{\partial B}{\partial x} = - \frac{\partial B}{\partial I(t)} \quad . \qquad (5)$$

However, since $I(t)$ is a time integral of $V(t)$ one may write:

$$\delta I = V(t).\delta t \quad . \qquad (6)$$

Inserting (2.6) in (2.5) leads to

$$\frac{\partial B}{\partial x} = \frac{-1}{V(t)} \frac{\partial B}{\partial t} \quad . \qquad (7)$$

Finally, rearranging egn. (2.7) one obtains the desired expression for instantaneous velocity:

$$V(t) = - \frac{\partial B/\partial t}{\partial B/\partial x} \quad . \qquad (8)$$

A simple implementation of this result for injected bolus experiments is to use the time density curve obtained at two sites "A" and "B" several mm apart. In this case one may write approximately:

$$V_{AB}(t_i) = -\frac{B_B(t_i) - B_B(t_{i-1}) + B_A(t_i) - B_A(t_{i-1})}{B_B(t_i) - B_A(t_i) + B_B(t_{i-1}) - B_A(t_{i-1})}, \quad (9)$$

where $B_A(t_i)$, $B_B(t_i)$ are the density value at sites A and B respectively in frame (time) t_i.

The instantaneous velocity subroutine used density data stored for the selected sites. The sites were grouped into equally separated pairs and the digitized density differences between sites and frames were computed according to eq. (9). Usually 4-5 pair were taken and the instantaneous velocity respresents an averaged value of those pairs.

RESULTS

C.O.G. Method

Results of the stationary flow study are shown in Figure 2. A good correlation is observed (correlation coefficient is .960). However, the results are quite scattered with a mean deviation of 24% over the whole range of velocities covered (9-80 cm/s). The method was checked by using different site-separations (4mm-3cm) and no significant changes were observed.

The pulsatile flow results are shown in the second column of Table 1. The following points should be noted:

(i) The pump cycle lasted about 49 frames whereas at average velocities of 36 and 31 cm/s the bolus lasted for a shorter period. The averaging over cycle was, therefore, incomplete.

(ii) It is observed that no substantial change is obtained by using wider site-separations.

(iii) The reduction in flow was accomplished by narrowing the inlet tube to the pump so that at velocities below 31 cm/s an increasing backflow component was visually observed as flow was further reduced. It is possible that backflow persisted also in the study where V = 36 cm/s.

As may be seen, there is no significant correlation between the known and calculated results using the C.O.G. method in this case.

"Instantaneous Velocity" Method

A typical result of a study under constant flow conditions performed at four (4) sites separated 4mm apart is shown in Table 2. As may be seen, the results at different sites in each frame are consistent. It is observed that the largest velocities were observed in the first few frames, followed by a general decrease with time.

Figure 3 shows the correlation obtained between the average instantaneous velocity computed from the cine-densitometric data and the experimentally determined average velocity in the constant (gravitational) velocity case. The correlation coefficient is 0.99 in this case. The mean deviation from the unity slope line is 8%. The error bars refer to mean deviations of the instantaneous velocity (non-averaged) in each study. The average deviation of those is 35%.

In Figure 4, a sequence of cine frames from a pulsatile flow study is shown; the bolus is seen to appear and progress in one direction (downstream) then a reversal of the direction of motion is observed. Table 3 shows the resultant instantaneous velocity at these sites, its average and standard deviation. It is observed that the positive values for the forward flow change into negative during backflow as expected.

Figure 5 shows the correlation between computed instantaneous velocities averaged over a cycle (810 msec) and the experimentally determined average velocity (measured over 70 cycles) in the pulsatile study. It is observed that the correlation is good (r=0.91). The standard deviation is 10%. At high flow rates, corresponding to 20cm/s, the persistence of the bolus at the sites studied was too short, thus the averaging was performed over about 80-90% only of the pump's cycle. At measured flow rates corresponding to 10,16, 20 cm/s, backflow was observed visually. Table 1 summarizes the results of the instantaneous velocity along with those of the C.O.G. method.

Figure 6 shows the results of similar flow measurements in a dog's descending aorta along with a typical density curve at one of the sites studied. The general features of forward flow, the peak value of forward flow,backward flow region and the one cardiac cycle separation between the peaks are in agreement with the experimentally determined velocity distributions. For comparison, an example of velocity distribution in the dog's femoral artery is shown in Figure 7. Similar results for two other dogs studied are shown in Figures 8 and 9.

DISCUSSION

Limitation of the C.O.G. method under pulsatile flow conditions

As evidenced by the experimental results, the traditional C.O.G. method cannot be used in high velocity pulsatile flow cases. In the following, the general limitation of the C.O.G. method when applied to pulsatile flow are discussed.

The first limitation of the application of the C.O.G. method to pulsatile flow condition is related to the time averaging period. The persistence of the bolus at the sites studied depends on such factors as the injection time, the phase at the cardiac cycle in which the injection is made and the magnitude of flow. The bolus length at the sites cannot therefore be "tailored" so that it will cover an exact integer number of cardiac cycle(s). However, in order to obtain the true mean flow under pulsatile condition, one must have an exact average over complete cardiac cycle(s). Neglect of this point may have been another cause for the widely scattered results obtained in the past.

The second limitation of the method is related to cases where a large back-flow (BF)component is present. In this case, one must reanalyze the concept of transit time, as the dye material moves back and forth between sites "A" and "B" (where "B" is located downstream with respect to "A".) As was pointed out above, the only meaningful discussion of average flow and of the corresponding transit time will be to analyze complete cardiac cycle(s). The following discussion deals therefore with transit time definition where one complete cycle is analyzed.

Consider first a single dye particle travelling back and forth under back-flow. It is clear that in this case the definition of average transit time between "A" and "B" should read: The time between the particle first arriving at "A", and its lastly leaving "B". When a continuous distribution of dye "particles" is present, one may define two distribution times: The time distribution of the particles' first appearance at "A" and the time distribution

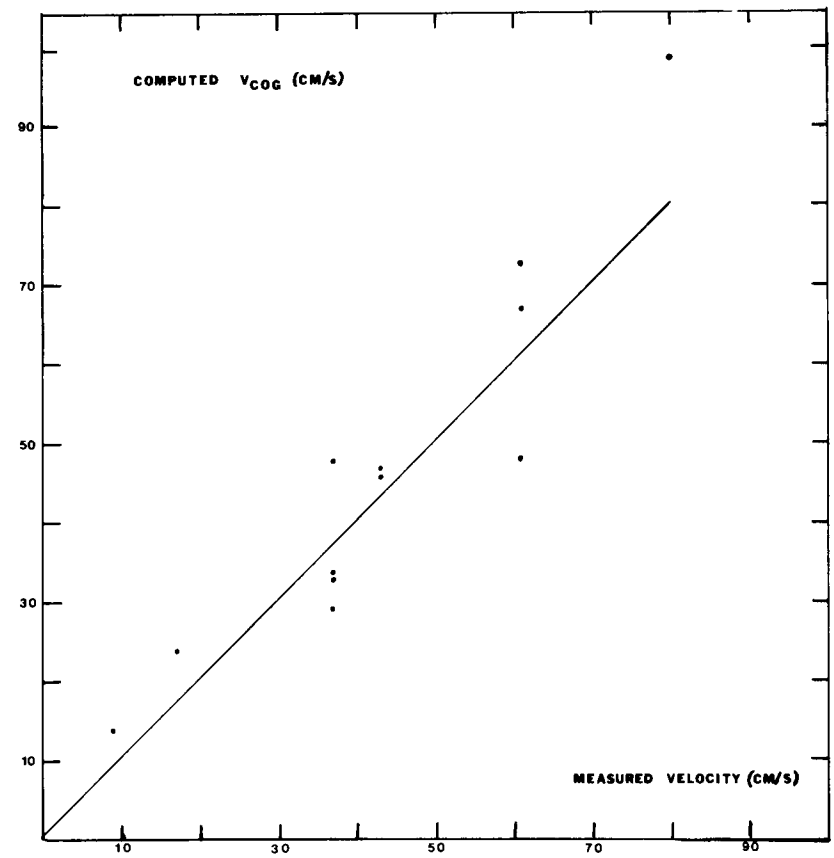


Figure 2. The correlation between computed and measured flow velocities using the C.O.G. method in a study at constant gravitational flow.

TABLE 1

PULSATILE FLOW

VELOCITIES OBTAINED BY THE C.O.G. AND

INSTANTANEOUS VELOCITY METHODS

Measured velocity cm/s	$\bar{V}$ C.O.G. cm/s	$\bar{V}$ Instantaneous cm/s
36	40;47;55	39*
31	67;31	28*
20	31;47	18;17*
16	65;61	17;17
11	53	10;11

* Averaged over .8-.9 of a cycle

TABLE 2

GRAVITATIONAL FLOW-INSTANTANEOUS VELOCITY DETERMINATION

V MEASURED = 43 cm/sec

DISTANCE BETWEEN SITES: 0.4 cm.

FRM	INSVEL1	INSVEL2	INSVEL3	INSVEL4	AVEL	STAN.DEV.
1	73.8	69.3	72.1	*124.1	84.8	22.7
2	67.6	54.4	67.2	83.7	68.2	15.5
3	50.3	39.4	31.9	*-3.7	29.4	20.2
4	10.7	10.9	23.2	28.5	18.3	7.7
5	70.0	39.6	50.7	30.8	47.8	14.7
6	47.0	55.9	*106.2	33.4	60.6	27.6
7	22.8	36.5	50.4	24.8	33.6	11.0
8	35.1	32.1	38.7	28.1	33.5	3.9
9	27.3	24.5	19.7	19.5	22.7	3.3
10	6.2	17.2	14.9	9.2	11.6	4.6
11	28.0	37.3	27.5	21.2	28.5	5.7
12	18.6	31.2	20.9	19.3	22.5	5.1
13	25.6	30.1	23.2	22.0	25.3	3.1
14	26.2	26.6	25.0	20.7	24.6	2.4

AVERAGE = 37 cm/sec

* High deviations

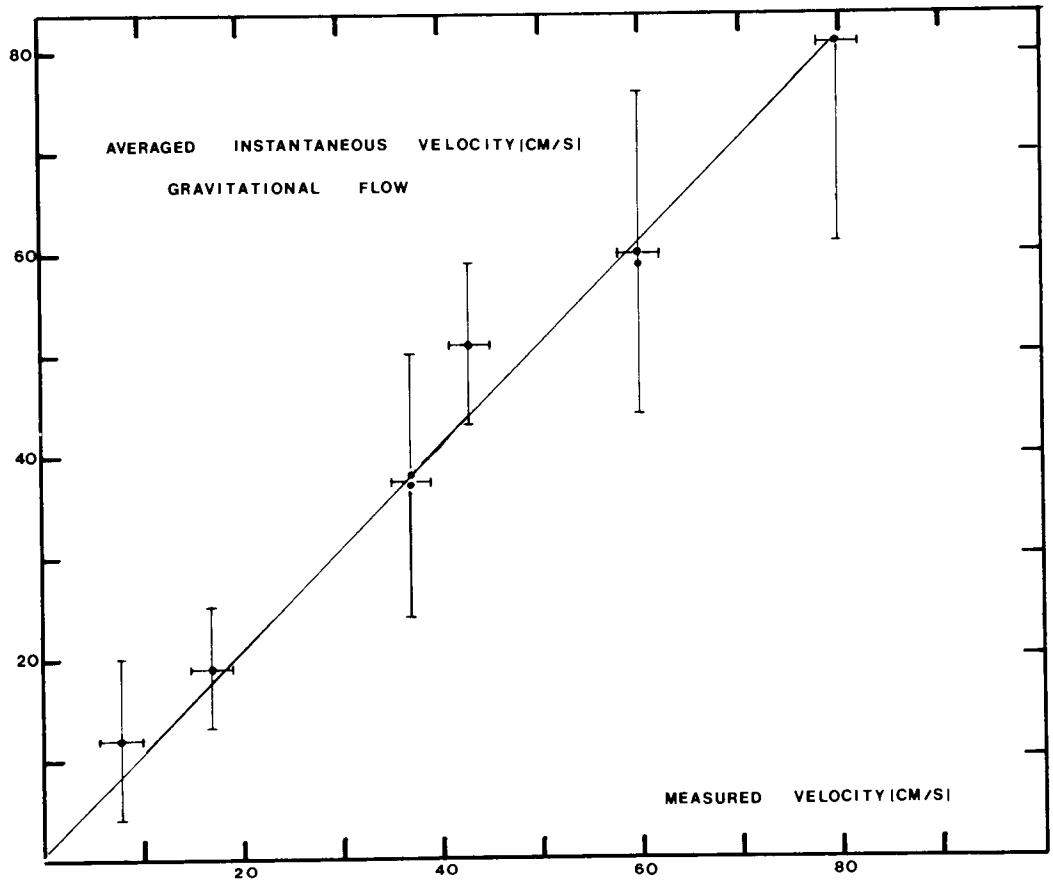


Figure 3. Results of the average velocity obtained using the instantaneous velocity method vs. measured flow velocity in the constant flow study.

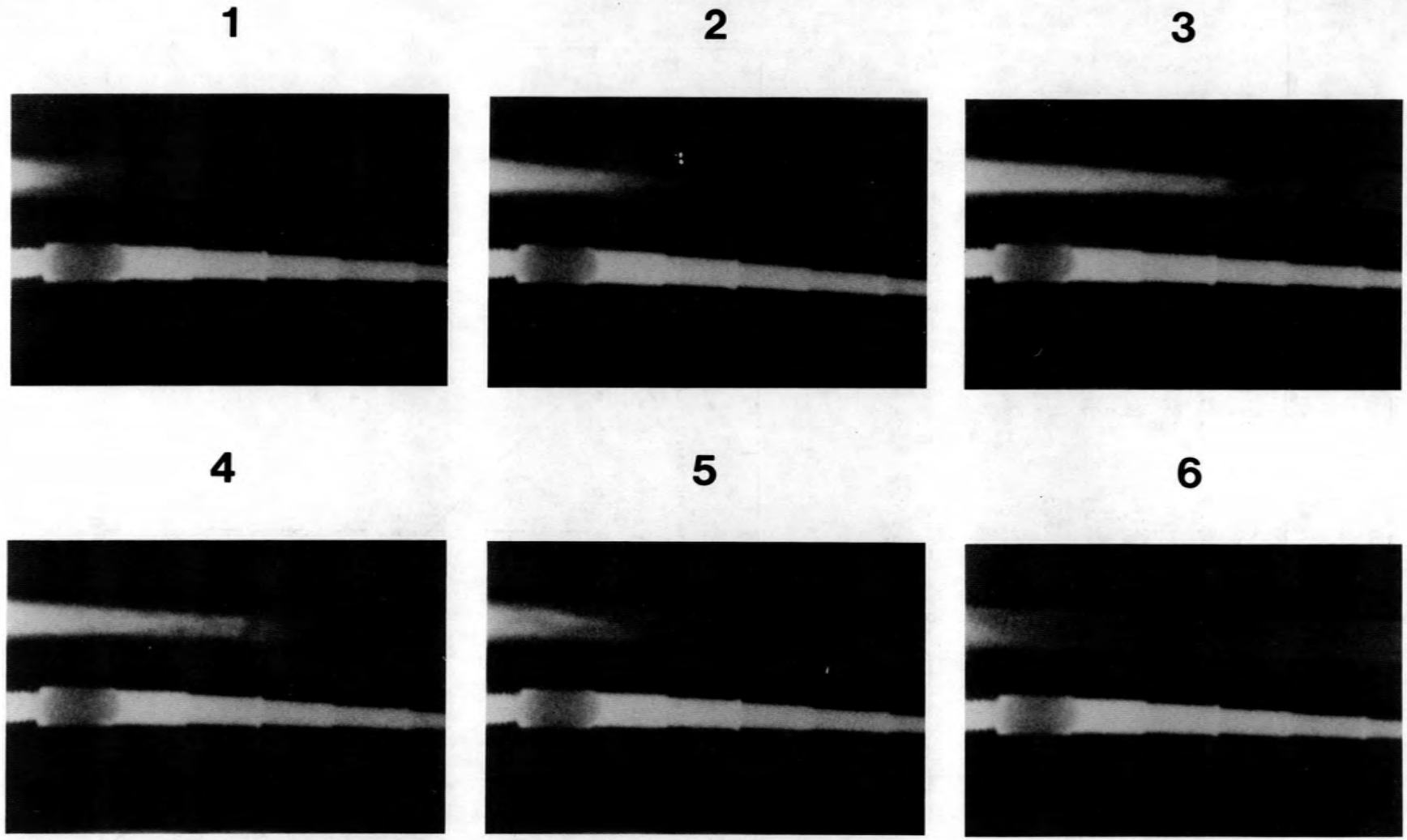


Figure 4. A sequence of 6 cine frames showing a case of backflow. The bolus enters and then reverses its direction clearing the region of vessel analyzed.

TABLE 3

INSTANTANEOUS VELOCITY RESULTS

FOR THE CASE OF BACKFLOW

(cm/s)

FRAME	V(1)	V(2)	V(3)	V(4)	V(5)	AVERAGE	STAN. DEV.
1	39.7	40.9	38.8	36.0	36.5	38.0	1.9
2	51.6	49.7	45.6	39.5	39.5	45.2	5.0
3	31.0	32.4	35.2	38.8	31.1	33.7	2.9
4	14.9	20.6	11.7	8.1	15.2	14.1	4.1
5	4.6	-10.1	-4.3	3.5	2.7	-0.7	5.7
6	-14.4	-8.9	-10.5	-14.2	-16.7	-12.9	2.9
7	-30.9	-31.3	-26.0	-19.9	-20.3	-25.7	4.9
8	-39.8	-39.5	-35.6	-31.8	-32.0	-35.8	3.5
9	-41.2	-43.6	-51.9	-61.6	-69.3	-53.5	10.7
10	-34.5	-36.2	-32.6	-35.5	-43.7	-36.5	3.8

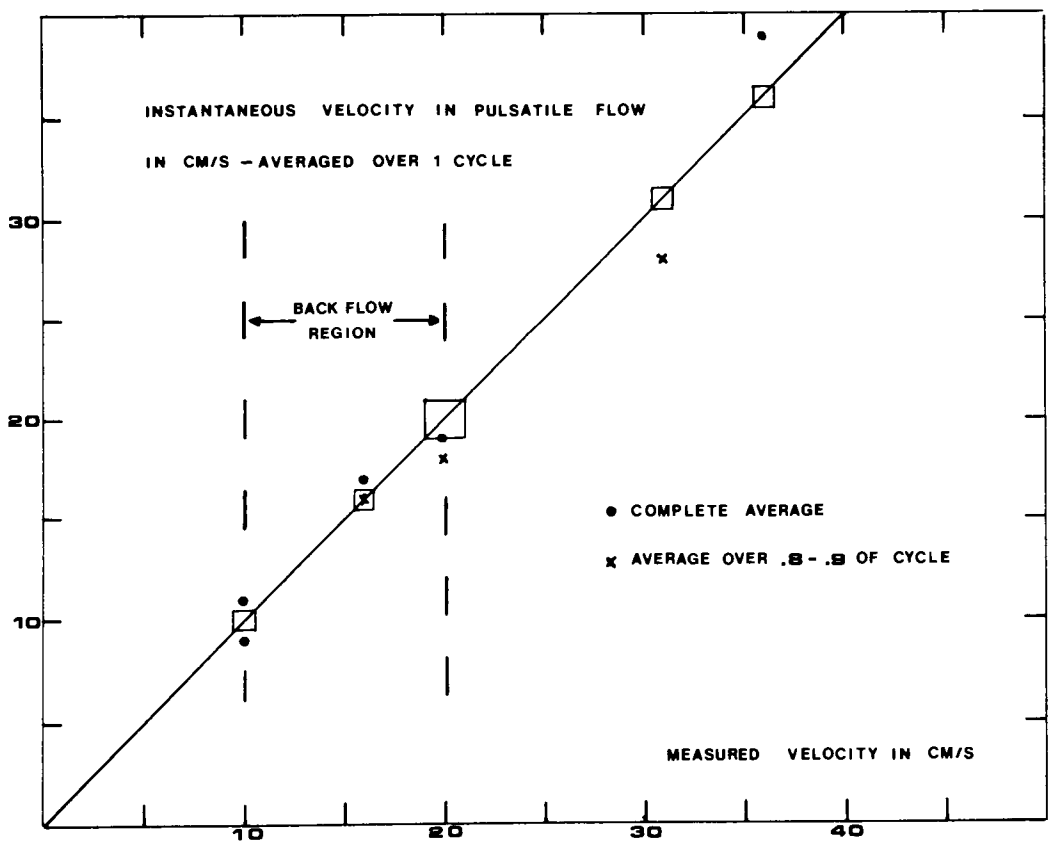


Figure 5. Results of instantaneous velocity averaged over one pump-cycle in a pulsatile flow study. Squares indicate the regions of uncertainty due to experimental errors.

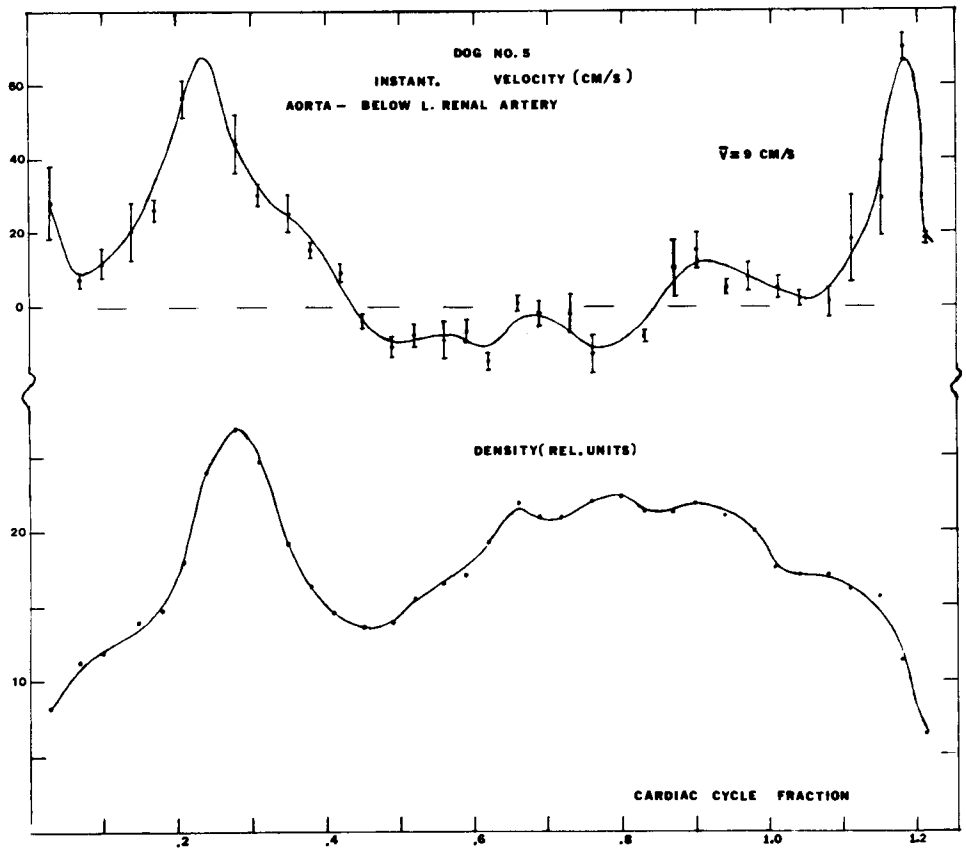


Figure 6. The application of the instantaneous velocity method to dog's arterial blood flow data taken below left renal artery. Top: Instantaneous velocity distribution. Bottom: Typical density curve at one of the sites analyzed.

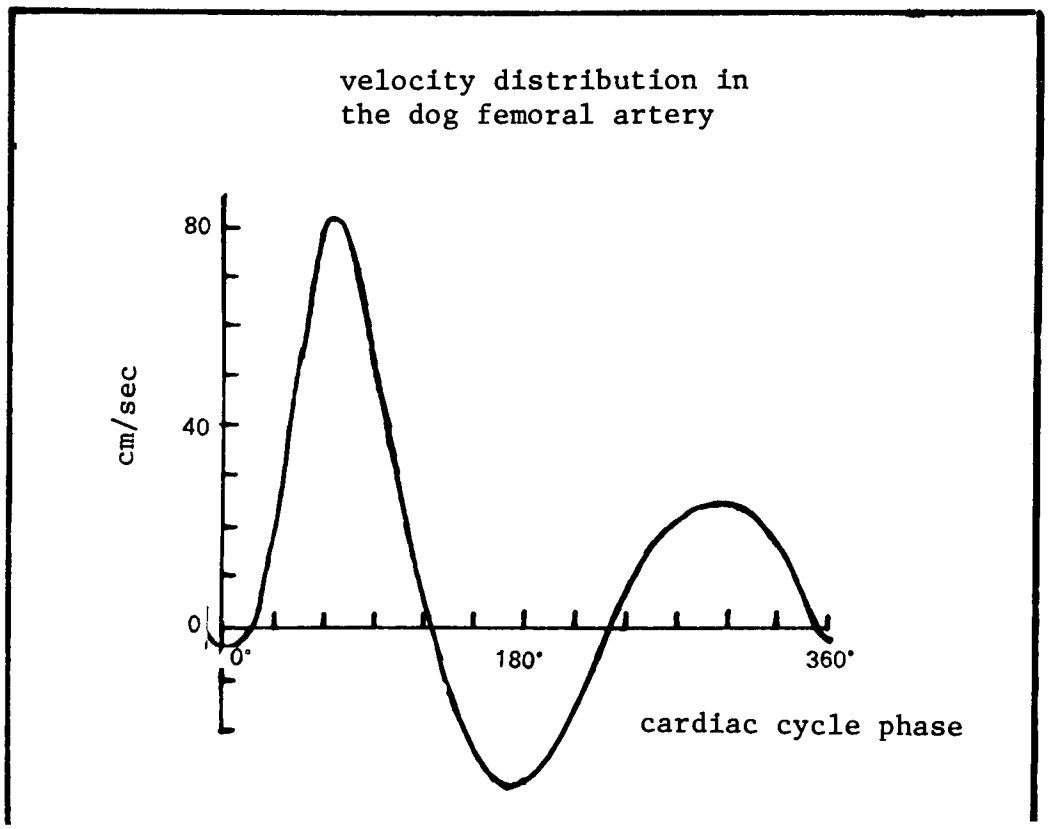


Figure 7. Velocity distribution in the dog's femoral artery-using high speed optical cineatography (5).

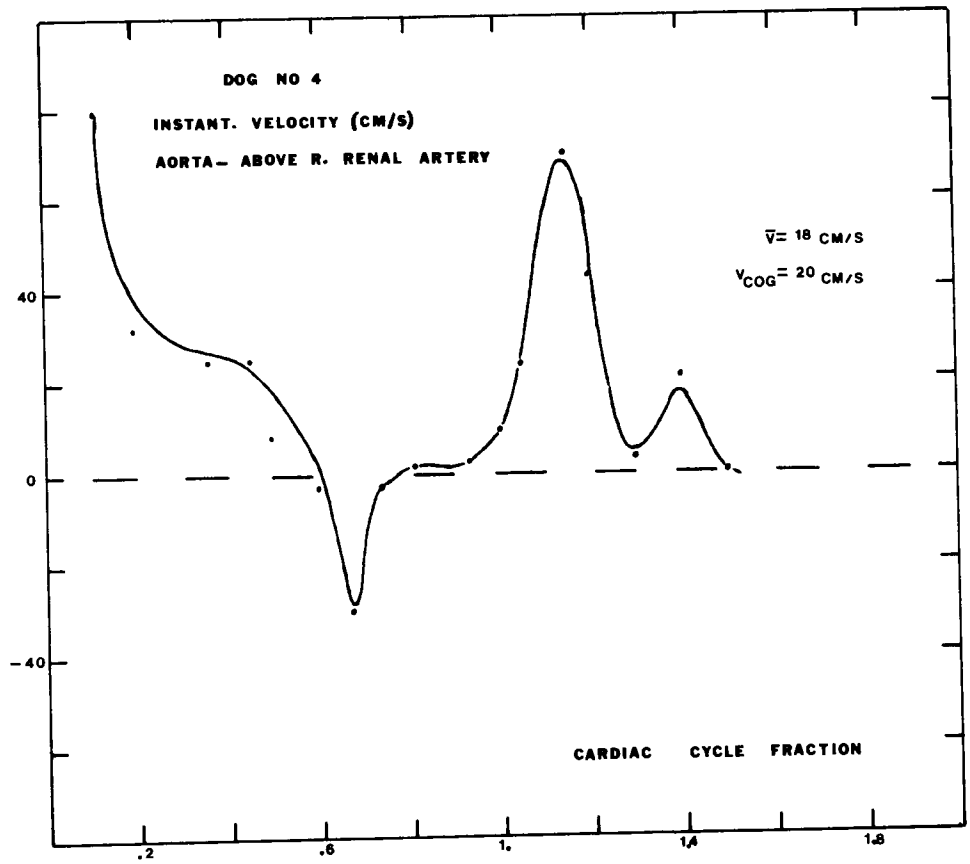


Figure 8. The instantaneous velocity distribution determined from cine data of dog's aortic blood flow.

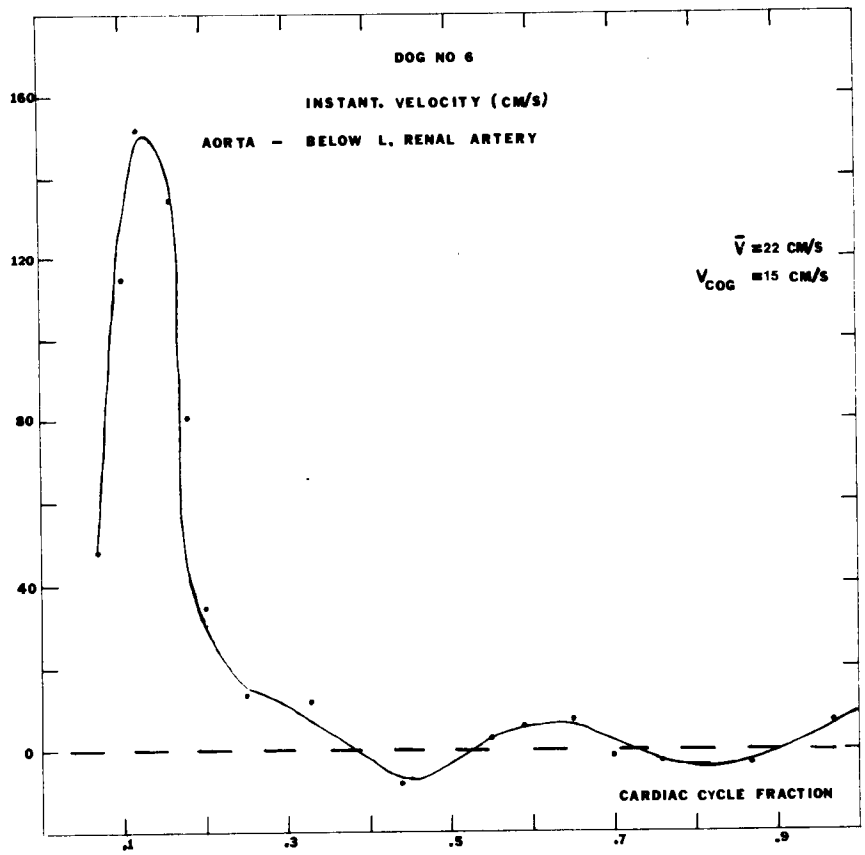


Figure 9. The instantaneous velocity distribution determined from cine data of dog's aortic blood flow.

of their finally leaving "A". The transit time in this case will be the time difference between the averages (or centers of gravity) of those two time distributions. However, these time distribution curves do not coincide with the density curves obtained by measuring the density vs. time of the dye in each site. This is so since while measuring density at the sites, each crossing and recrossing of the dye material at the sites will appear as peaks in the density curves, whereas by the definition of the time distributions stated above, only one passage through each of the "A" or "B" sites will be recorded in this case--namely, the first crossing of the "A" site and the last crossing of the "B" site.

It should be noted that some error is also expected when calculating the transit-time using the cross correlation (6) method. This is so, since this method correlates the shapes of the density curves at the two sites to obtain the mean time difference between them. The occurance of several peaks at site "A" and the corresponding appearance of several peaks in the reverse (time) order at "B" due to the bolus recrossing of the sites under backflow, will introduce an error in the cross correlation and consequently in the transit time obtained.

One could overcome the first (time) limitation discussed above. In the case of an arbitrary bolus length, by computing the C.O.G. of those portions of the density curves corresponding to complete cardiac cycles. However, much of the information related to the transit time is contained in the time difference of the leading and trailing edges of the density curves. Thus, when using only a portion of the curve either the leading or trailing edge information would be lost.

A way to overcome the second difficulty is to use the C.O.G. method restricted to either the FF or BF regions(7). One may use multiple short bolus injections in various phases of the cardiac cycle to scan the whole FF or BF region. This method requires, however, that one be able to distinguish between the backward and forward flow regions. Difficulties are also encountered in the intermediate region in which the flow changes its direction.

The "Instantaneous Velocity" Method

The results obtained with the instantaneous velocity method are reasonable for steady and pulsatile flow measurements in simulators and in real studies. The results are dispersed, due to noisy data, and hence signal averaging is required. Standard deviations obtained at the constant flow experiments are about 20-30% for the instantaneous (non-average) values. Higher deviations are observed in the pulsatile flow studies. There are several causes for these fluctuations:

(i) As previously stated, we have taken the simplest way to implement the instantaneous velocity relations (8) by computation based on (9). This simple form is not suitable for use in cases of either high velocities or highly non-uniform bolus profile along the vessel. In both cases, the temporal and spatial derivative will show large variation. To overcome this difficulty one must use a program which will select the appropriate site and temporal separation based on initial velocity information obtained e.g. by using the present method of instantaneous velocity. This will then be repeated iteratively to minimize deviations.

(ii) Due to noise present in the digitizing system, spatial and temporal gradients below a certain level will have to be rejected. In addition, a smoothing of the density curves must be applied. This may involve either partial integration over several frames or the use of low pass Fourier Filtering.

(iii) As indicated previously, there is a general tendency for instantaneous velocity values to decrease with time in the constant velocity studies. This is explained as follows: Since the velocity profile across the vessel is close to parabolic, the bolus will first appear at the central (high velocity) portion of the sites. This central region will gradually be depleted and the bolus's "tail" will be confined mainly to the lateral (low velocity portion of the vessel). This phenomenon is shown in Figure 10 where the density profile across the vessel at a certain site is shown. A high central density is observed at the first few frames (1-6) a "normal" shape at frames (8-10) and a center depleted density at the end frames (16-24). To overcome this difficulty one must use the density profile across the vessel and the experimentally determined parabolic velocity profile constants. These may be determined in one measurement as they are not expected to be velocity dependent.

It should be noted that whereas this phenomenon does affect the instantaneous velocity values, it tends to be averaged out when average velocity over cycle(s) is computed.

An additional advantage of this "differential" method over the C.O.G. "Integral" method is its relative insensitivity to non-linearities present in the imaging or recording systems. This is so, since in analyzing density curves using the C.O.G. method, one must use the whole dynamic range of the images. Densitometric results will therefore be affected by the non-linearities present in either the imaging system (image intensifier) or recording media (film). When using the instantaneous velocity method, one deals with differences of images which is equivalent to using a small dynamic range of the image.

CONCLUSIONS

An instantaneous velocity method for the determination of both instantaneous and average blood flow has been developed for cine-or video-densitometric angiographic data. The method has been validated under stationary and pulsatile flow condition. The velocity distribution obtained by applying this method to experimental animal studies are in agreement with the expected results.

The main advantages of this method are:

1. Instantaneous flow information is provided.

2. It permits the determination of average flow under highly pulsatile conditions.

3. Insensitivity to non-linearities in the imaging or recording systems.

4. The use of comparatively small (~1cc) amounts of contrast dye.

The drawbacks in this method are:

1. Highly fluctuating results and the consequent need for a smoothing procedure.

2. Some correction must be applied to instantaneous velocity data if velocity distribution is needed.

The limitations of the C.O.G. method under pulsatile flow conditions have been discussed. This method is applicable under two main restrictions:

1. An average over complete cycle(s) must be made.

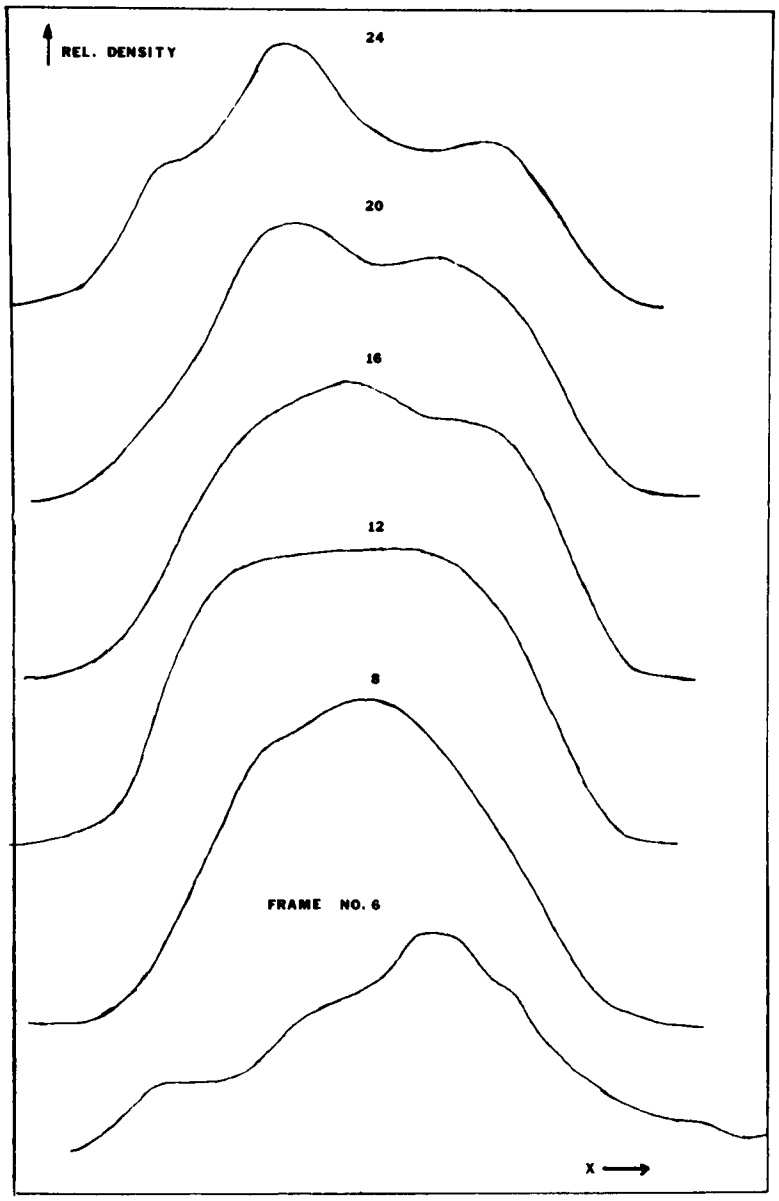


Figure 10. Density profile across the vessel taken at the same site during the various phases of the bolus progress. Frame #6-appearance of the bolus at the site, frame #24-low density "tail" of bolus is shown.

2. Low component of backflow, i.e. the time integral of the negative flow region must be small compared to the whole integrated flow.

Acknowledgement

The authors gratefully acknowledge the assistance of Mr. G. Novak and Mr. J. Norris in conducting the flow experiments.

REFERENCES

1. Rutischauser, W., et al.: Am. J. Roentgenol. Rad. Therapy and Nuc. Med. 109, 12, (1970).

2. Rutischauser, W., et al.: Circulation 36, 961, (1967).

3. Itzchak, Y., et al.: Invest. Rad. 10, 1 (1975).

4. Smith, H.C., et al.: "Simultaneous Indicator dilution--curves at selected sites in the coronary artery saphenous vein grafts by Roentgen Video-densitometry," Roentgen, Cine and Videodensitometry. Ed. P.H. Heintzen Stuttgart, Thieme, 1971, pp. 152-157.

5. McDonald, D.A., et al.: Blood flow in arteries, Ed. Williams and Wilkins Co. 1974, p. 130.

6. Rosen, L., et al.: Radiology 103, 305 (1973).
Silverman, N.R., et al.: Invest. Radiol. 12, 319 (1977).

7. Kedem, D., et al.: "Velocity distribution and blood flow measurements using video densitometric methods," Invest. Radiology - to be published.

DISCUSSION:

In reply to a question from K. Britton, Effron stated that in calculating flow, blood vessel cross sections were assumed to be circular. B. Oppenheim asked why corrections for the flow profile across a vessel are required in determining transit time: do not flow measurement determinations of transit time inherently average over the various simultaneous velocities? Effron explained that corrections for backward flow components were required in order to separate forward flow from net flow.

TOPICS IN ULTRASOUND IMAGE PROCESSING

A PANEL DISCUSSION

Panel Members:
J. Greenleaf, Mayo Clinic, Rochester, Minnesota
B. R. Pullan, University of Manchester, England
Chairman:
J. Goddard, Vanderbilt University Medical Center
Nashville, Tennessee

Recent Developments in Ultrasonic Tissue Characterization

The first topic was introduced by J. Goddard who reviewed some of the important developments at the recent Second International Symposium on Ultrasound Tissue Characterization (N.B.S., Gaithersburg, Maryland, June 13-15, 1977). The symposium consisted of papers on (a) techniques for measurement of basic ultrasonic parameters, (b) signal processing and pattern recognition techniques, and (c) ultrasonic properties of tissue and their dependence on physical and biological variables. He noted that one of the strong themes of the symposium was the presentation of the latest work in the effort to detect the early stages of breast cancer by ultrasound. Two papers on this subject were considered to be outstanding and were briefly reviewed. In one of the papers, "Tumor Detection by Ultrasonic Doppler Blood-Flow Signals" by P.N.T. Wells et al, some preliminary results of Doppler blood flow measurements in normal and tumorous breasts were described. The authors stated that they were able to differentiate normal and tumorous breasts by detecting the elevated blood flow associated with the increased vascularity of the tumor.

In a second paper, "In-Vivo Measurement of Ultrasonic Refractive Index Distributions in Human Breasts with Time-of-Flight Tomography" by G.H. Glover, a time-of-flight, through transmission tomograph system, was described. The author had used the system to differentiate between malignant and benign masses in the breast on the basis of the deviations in acoustic speed. J. Greenleaf indicated that his group was involved in a similar type of ultrasonic computer assisted tomography of the breast and went on to describe some of G.H. Glovers results in more detail. He pointed out that G.H. Glover had made a histogram of the acoustic speeds at each pixel in a plane through the breast and found that the histogram shifted to higher velocities in tumorous breasts. He stated that this also occurred in planes in the tumorous breast which did not contain the cancer. J. Greenleaf added that he had observed the same phenomenon in his own work on excised breasts.

J. Goddard remarked on the current controversy about the hazards of X-ray mammography and wondered whether any comparative studies had been performed to determine if ultrasound would be a suitable alternative to X-rays in the diagnosis of breast cancer. In response, J. Greenleaf stated that his own work was in an early phase and that he had not yet had the opportunity to carry out a comparative study. B. Pullan queried, that

since ultrasound basically measured the same parameter as CT scanning, namely changes in density, whether ultrasound devices would be any better than CT scanning in detecting breast cancer. Disagreeing, J. Greenleaf replied that in ultrasound one is basically measuring acoustic speed which is a function of the stiffness and density of the tissue. Furthermore, that the changes in density are small compared to the changes in stiffness which is the reason that large breast tumors can be located by palpation. He added that the interaction of acoustic energy with a tumor was very much stronger compared to X-rays, and that so far ultrasound had been moderately successful in the detection of breast tumors. B. Pullan then asked, in view of the strong interactions of ultrasound with tumors, why the present generation of ultrasound scanners had not been much more successful in this area. D. Krause felt the reason for this was due to the relatively poorer resolution of ultrasound devices compared to X-ray scanners.

D. Krause queried the role of spectral analysis measurements in tissue characterization. In reply, J. Greenleaf pointed out that the effects of scatter on acoustic energy were very frequency dependent because certain frequencies were resonant to the periodic structure of the tissue under study. Fourier analysis of back scattered energy could, in principle, yield some information about the distribution of sizes of, for example, blobs of fat which were suspended in the tissue. Two types of scatter measurements could be performed, one as a function of frequency at a fixed angle, the other as a function of angle at a fixed frequency. In addition, Greenleaf stated that the frequency content of back scattered energy was a strong function of geometry because parallel planes of tissues existed which act as phase filters, producing resonances at certain frequencies. When ultrasound was reflected from such tissues the resonant frequencies would therefore either be very high or low in amplitude. He added that tissue attenuation and absorption were strongly frequency dependent and so spectral analysis might also yield some information about such parameters as well.

Archiving of Medical Images

B. Pullan introduced the topic by describing some of his experiences with whole body CT scanning. He stated that in a typical patient study ten pictures consisting of 320 x 320 elements were collected. The data was archived on magnetic tape and even with only a small patient throughput, a 2400 foot reel of tape was filled each day. This method of archiving often resulted in time-consuming delays when an individual patient record needed to be retrieved at a later date. He added that he felt similar inconveniences would also be encountered in other imaging modalities such as phased array ultrasound which produced data typically consisting of 200 x 200 x 100 elements and in nuclear magnetic resonance. In the latter, genuine three-dimensional data of great biological significance was being generated which consisted typically of 160 x 160 x 160 elements.
B. Pullan doubted that fast access bulk storage for archiving medical images could be economically realized with present technology and queried whether data compression techniques might be an alternative. He then asked the audience for their advice and possible solutions to the problem.

Ashare responded that he had been working with data compression techniques and had obtained a data reduction of a factor of four for nuclear medicine images. He felt that a factor of ten should be readily achievable without sacrificing image quality.

D. Kirch considered the only viable bulk storage facility at present to be 9-track magnetic tape. He suggested that, as an intermediate bulk storage facility, a 330 megabyte disc could be used. This would be sufficient, for example, to accomodate one month's work of CT, nuclear medicine and ultrasound scans.

S. Pizer commented that in his experience he had found that, compared to magnetic tapes, floppy discs offered much faster access times, were a great deal more portable and were much easier to file.

D. Janney pointed out that the problem was not too dissimilar to that of land-satellite transmissions where hundreds of photographs of the earth required archiving. In that particular application a large bulk storage device holding, say, 3 billion words of data and costing, say, $1,000,000 was used. He queried whether such a device would be feasible for medical image archiving.

B. Pullan replied that the cost was likely to be prohibitive considering that a device would be required in every major hospital.

A. Brill discussed the possible application of a video disc system. He commented that his group had had five years experience with the last generation of video-disc systems but had found that the performance was disappointing. However, he believed the quality of the current generation of video discs to be considerably better. A typical system was able to store a large amound of data in a compact manner; for example, one disc could typically hold 30 minutes worth of data packed at the rate of 30 frames per second. He stated that the Phillips company in Europe was developing such a system for image archiving but that these systems provided only marginal quality and were still rather expensive.

B. Pullan thought that, considering the enormous dynamic range of NMR, CT and ultrasound images, the data would have to be stored in digital format. He believed that the bulk and the slow access times of magnetic tape would be prohibitive and suggested that new technologies might provide better alternatives in the future. D. Ochs indicated that one of the promising areas in this regard was the development of optical computing and storage techniques. S. Pizer agreed and added that magnetic bubble memories might also find application. He thought that technology would in the near future be producing mass memories whose price would not be restrictive.

J. GREENLEAF: "ANALYSIS OF TRANSMISSION ULTRASOUND"

J. Goddard asked if any allowance had been made in the reconstruction algorithms for tissue anisotrophy. In reply, J. Greenleaf said that no allowance was currently being made, but that he was studying the problem. He added that the present vector reconstruction method essentially came about by attempting to develop an anisotrophy reconstruction method.

An unidentified member of the audience queried how many views and what type of method was used in the image reconstruction. J. Greenleaf answered that a convolution type method based on 30 views was used. He commented that the gain was very high so that one is looking at very small alterations.

A. Todd-Pokropek inquired to what degree the results obtained from excised samples were distorted compared to the in-vivo results. J. Greenleaf replied that his group had studied some 80 excised breast samples but that the results would not be reported until they had been fully correlated with in-vivo measurements. He stated that, so far, the in-vivo results were the same as the excised except for the fat component, which showed a reduction in acoustic velocity.

J. Goddard asked about the feasibility of improving the image quality through the use of acoustic contrast agents. He pointed out that in the session, two agents, namely ringers solution and heat, had been mentioned as producing changes in acoustic velocity. J. Greenleaf responded that he had studied the effects of RENOVIST, which has a high density and found that it produced a change in acoustic speed. He added that indocyanine-green dye is a very good ultrasound constrast material, possibly because it produces micelles when it is first injected which then subsequently disintegrate. Indocyanine green dye is not a contrast material as soon as it passes through a capillary bed, so it is one of the very interesting contrast agents which can tag the venous phase but will not show up in the arterial phase. He then stated that he was currently searching for a tracer which would tag only the arterial phase, so that following an intravenous injection, both venous and arterial phases could be studied independently. J. Goddard then inquired whether acoustic contrast agents could be used to make quantitative estimates of flow in the body. J. Greenleaf replied that he thought this would be difficult because the reflected intensities were not linearly dependent on tracer concentration, but added that a linearization technique might be a possible solution.

HIGH-SPEED ULTRASONIC SCANNERS

W. A. Anderson
Systems & Techniques Laboratory
Varian Associates, Inc.
Palo Alto, California, U.S.A.

ABSTRACT

High-speed ultrasonic scanners are defined as two-dimensional imaging instruments that can produce cross-sectional ultrasonic images at a rate of at least 10 per second. Basically there are two approaches to high-speed scanning. These are mechanical scanning systems and electrical scanning systems. Within each class there are several different approaches and each of these will be described with its advantages and limitations. An electrical scanning system using a phased-array system is described in greater detail. The application of high-speed ultrasonic scanning to cardiac imaging is discussed. Information processing steps to yield quantitative data about cardiac performance are described and illustrated.

INTRODUCTION

The development of pulsed ultrasound for imaging has its origins in the work of Dr. F. A. Firestone at the University of Michigan.[1] Dr. Firestone developed a nondestructive testing technique by observing the reflected ultrasound signals. A short radio-frequency pulse was applied to a transducer causing it to emit a pulse of sound energy into the object being examined. Reflections from various discontinuities within the object were detected by the same transducer and displayed on the vertical axis of an oscilloscope. A linear scan voltage initiated at the same time as the transmitted radio-frequency pulse was applied to the horizontal axis. Knowing the approximate sound velocity within the object permitted the position along the horizontal axis to be related to depth. This method, called an "A-mode" scan, is illustrated in the upper part of Fig. 1a.

Although the work of Firestone was directed primarily towards nondestructive testing, it stimulated interest in the use of ultrasound for diagnostic imaging. By the late 40's and early 50's numerous investigators were using ultrasound for examining various body organs. Drs. Edler et al used ultrasound for monitoring the continuous movements of the heart.[2] The technique is called echocardiography. Typically the one-dimensional reflection image is displayed by brightening the spot on an oscilloscope as reflections are received. The horizontal axis corresponds to depth. A slow scan on the vertical axis permits

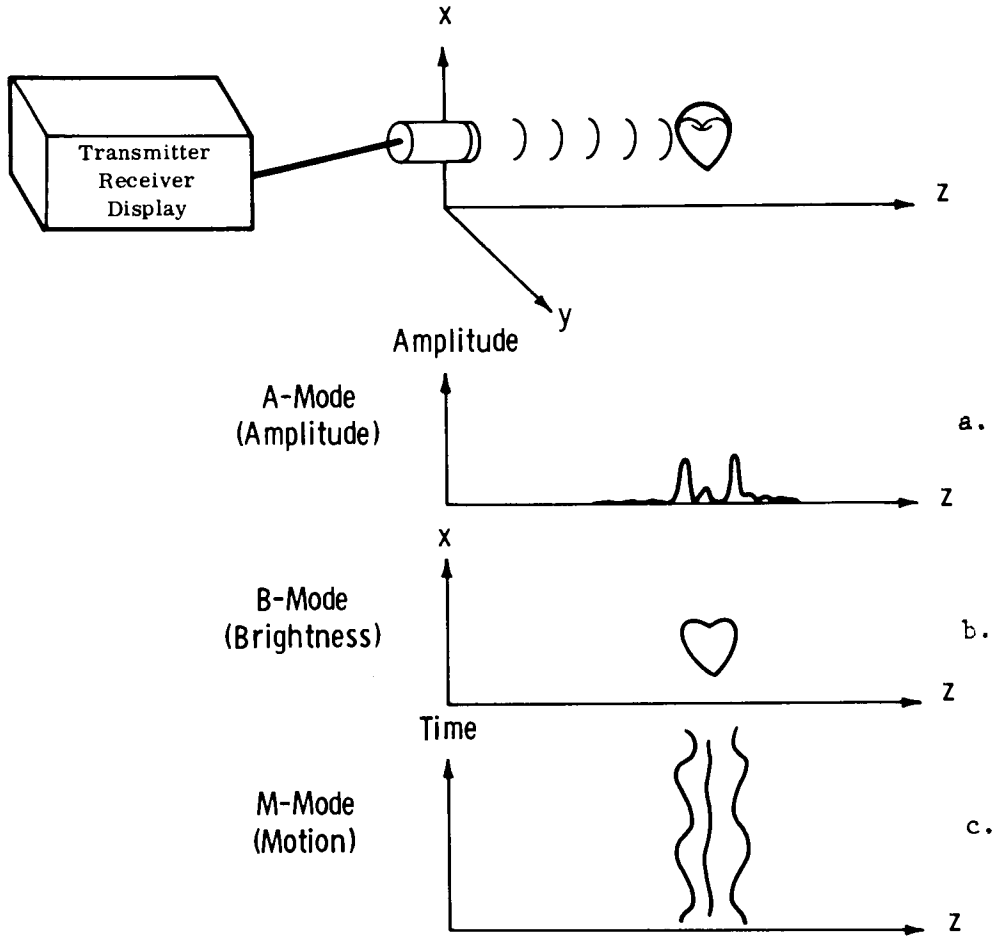


Fig. 1. The Three Ultrasonic Scanning Modes.

one to observe the cardiac motions through many cardiac cycles. This form of display is commonly called an "M-mode" (for motion) or "T-M mode" (time-motion) scan.

In order to obtain a two-dimensional cross-sectional image, a means of sweeping the ultrasonic beam through space is required. Various mechanical and electrical methods of achieving this sweep at high speeds are described in the next section of this paper. The display of echo image data in a two-dimensional format is called a B-mode (brightness) scan. Here one axis of the display corresponds to depth and the other to a transverse dimension. The three scan modes are illustrated in Fig. 1.

HIGH-SPEED SCANNING SYSTEMS

High-speed scanning incorporates the idea of generating pictures at a sufficiently high rate to avoid blurring from motions of the object being observed. To "stop" the motions of the heart wall requires that a picture be captured in approximately 0.1 second or less and considerably shorter times are required to capture some of the valve motions. To change the imaged area the operator must move the transducer system over the skin of the patient. Frame rates greater than 10 images per second provide the operator with an apparently continuously moving image that follows the transducer movement. For these reasons it is

important to have frame rates greater than ten images per second.

In order to achieve high-speed scanning one must be able to translate the beam quickly over the object being examined. The two basic approaches to achieve such a rapid scan can be grouped into mechanical systems and electrical systems. The mechanical systems may change the orientation of the transducer or of an intermediate element placed in the sound beam. Generally, electrical systems employ a multi-element transducer in which the excitation pattern is changed in order to steer and sometimes focus the sound beam.

Several approaches to high-speed mechanical scanning are illustrated in Fig. 2. A rocking transducer, illustrated in Fig. 2A, is achieved by the mechanical coupling of the transducer element to a suitable motor drive. A recent example of such a system has been given by Griffith and Henry.[3] Another mechanical approach, described by Holm[4], is illustrated in Fig. 2B in which several transducers are fixed to the periphery of a wheel which then can maintain sliding contact with the skin of the patient. By placing either one of these systems into a water bag, one can mechanically decouple the motion of the transducer from the patient as is illustrated in Fig. 2C. An early system of this type has been described by Wild and Reid.[5]

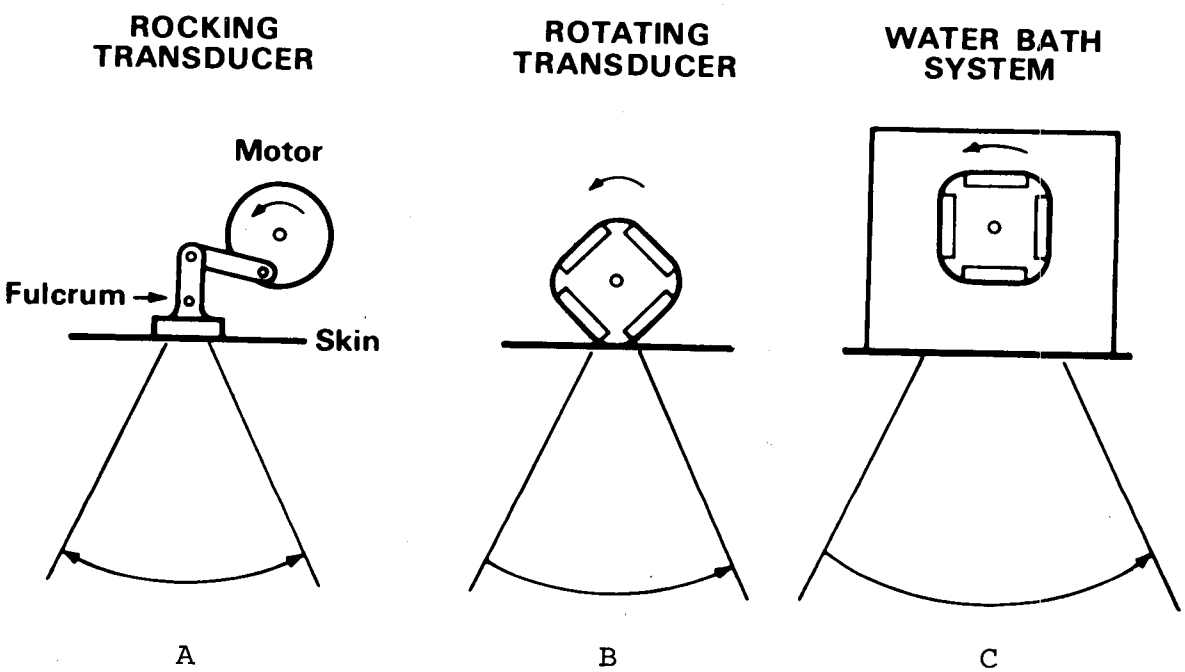


Fig. 2. Mechanical Scanning Methods

Three different approaches to electronic scanning are illustrated in Fig. 3. The system illustrated in Fig. 3A consists of an acoustical lens which focuses the acoustical image of an object on a plane containing a linear array of transducer elements. By activating a single element, the sound beam is emitted in the proper direction and, if desired, can be focused to a given depth. Returning sound waves from the same direction are intercepted by the lens and focused back onto the initial transducer element. The lens and linear array are placed in a water bath to provide sound coupling. Coupling to the skin is provided by a flexible membrane. By suitable multiplexing of the transducer elements, an array

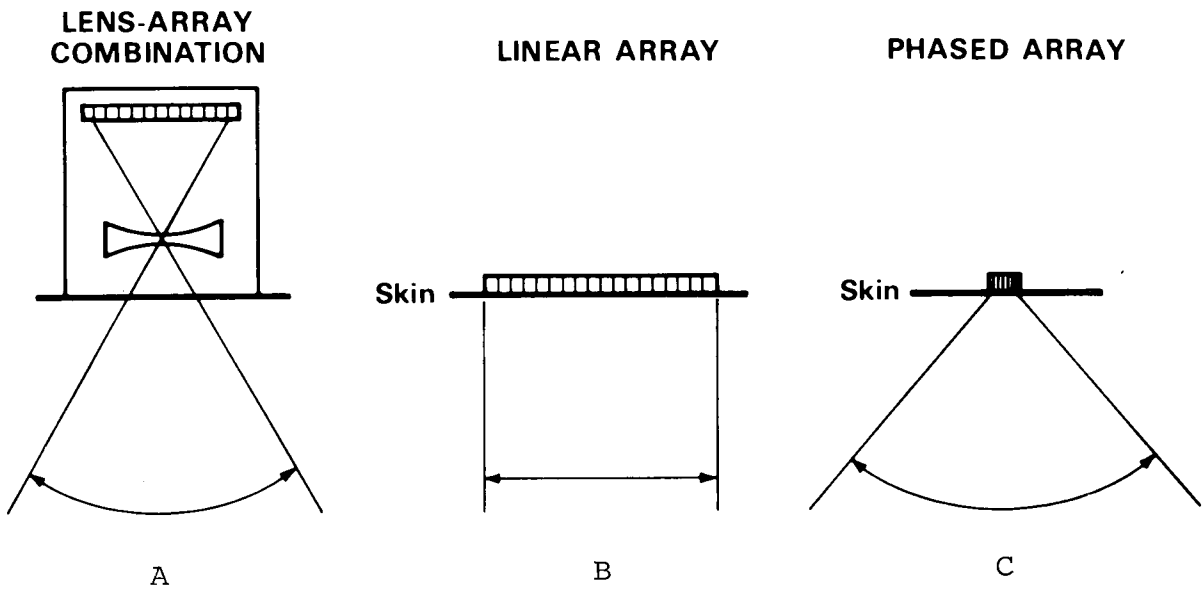


Fig. 3. Electrical Scanning Methods

of 2-dimensional picture elements can be built up on a cathode ray oscilloscope permitting one to obtain high-speed B-scan pictures.[6]

Another approach to electronic scanning is illustrated in Fig. 3B in which a linear array of transducers is again sequentially activated.[7] Here one or several adjacent transducer elements are activated to send out a sound wave that travels normal to the transducer face. The activated region is many wavelengths across so that in the near field the sound beam has the approximate size of the activated region. Any reflected signals are detected by the same transducer elements. After all of the reflected signals up to the maximum depth have been received, an adjacent group of transducer elements are activated. By continuing this sequence, a picture in a rectangular region is developed.

A third approach to electronic scanning is the phased-array approach[8] illustrated in Fig. 3C. Here the excitation of the individual elements is timed such as to produce a wave that is steered in the desired direction. In addition, focusing can be achieved by adding a quadratic delay component so that the beam will be focused at any desired depth. The same array of transducer elements is used to receive a wave from any desired direction by delaying the incoming signals appropriately before combining and detecting. Figure 4 illustrates how the waves from the individual transducer elements are summed to produce a resultant wave propagating in the desired direction.

Some of the advantages and disadvantages of these various high-speed scanning techniques are now considered. The electrically-scanned systems have the advantage of being able to rapidly switch scan direction so that one can maintain complete flexibility in the way that the images are built up. One consequence of this flexibility is that one is able to multiplex an M-mode scan with a B-mode scan and maintain a very high sampling rate for the M-mode recording. By observing the B-mode picture, the operator can find and maintain the desired transducer orientation while an M-mode scan is being recorded. Another advantage

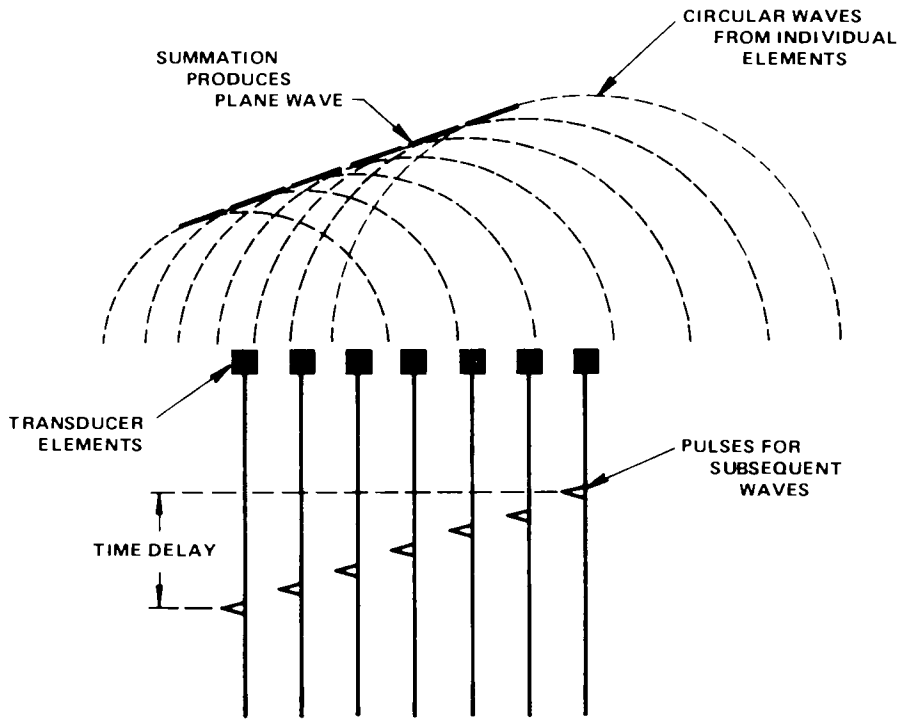


Fig. 4. Summation of Waves from Transducer Elements

of the electronic systems is their freedom from vibration, rubbing contact, or noise. The water bath systems have less vibration and noise; however, they have the disadvantage of requiring a support mechanism because they are generally too heavy to be held by the operator. The phased-array system presents the operator with the smallest, lightest-weight transducer for easy manipulation.

One problem in cardiac imaging is the limited accessibility of the heart, surrounded by the rib cage on the front and the lungs on the back. Thus for cardiac scanning it is desirable to use a sector scanner that permits the sound to be transmitted through a small intercostal space yet with the ability to scan over a large interior area. On the other hand it is this very feature that prevents the sector scanners from obtaining a large view of objects near the surface of the skin. The phased-array does have the advantage, however, of obtaining a large sector angle since the transducer itself can be maintained in flat contact with the skin while objects at an angle of 45° or more from the normal can be viewed. The mechanical systems have a cost advantage in that they require only one channel to transmit, receive and process the information. The linear array does not require the programmed time delays of the phased array. It can be multiplexed since only one element or group of elements are active at the same time. The phased-array system is the most complex electronically in that it requires one receiver channel for each element of the array as well as variable delays between the channels. This results in higher system costs.

APPLICATIONS OF HIGH-SPEED SCANNING

Perhaps one of the most obvious uses for high-speed ultrasonic scanning is to observe and follow the motions of a living heart. By clear visualization of the heart and its various parts, a number of cardiac abnormalities can be identified and to some degree quantified.[9] Examples of abnormalities that have been

visualized include prolapsing valves,[10] congenital and septal defects,[11] pericardial effusion,[12] stenotic valves and tumors or vegetation within the heart, the syndrome of IHSS[13] and conotruncal abnormalities.[14] Additional information obtained by various outlining or processing techniques is described below.

Another feature of high-speed scanning is that it permits a large region of the body to be studied cursorily, or a small region to be studied intensively in a short period of time. This makes it possible to increase the number of patients that can be seen per day; that is, the instrument is characterized by having a large "throughput." The fact that the image is viewed continuously as the transducer probe is moved permits extended objects to be followed and their three-dimensional structure reconstructed with little difficulty. It also makes it possible to locate small objects by quickly scanning through the region in which they are located.[15] For example ovaries, which are often difficult to locate by manual B-scanning systems, are readily seen in screen examinations of normal women with high-speed ultrasound scanning.[16] Another advantage of high-speed scanning is that it obviates the need for the patient to remain immobile or suspend respiration during the study. As a result infants, children, and comatose or uncooperative adults may be studied with ease. Another application of high-speed ultrasound scanning includes the ability to detect cardiac motion and "breathing motions" that occur in the fetus *in utero*. Verifying the absence of these motions provides a means of confirming fetal death *in utero*.[17]

With this introduction into high-speed scanning techniques we next turn to two information processing techniques that are presently being studied.

Light-Pen Measurement System

A light-pen measuring system is being designed to allow quantitative analysis of cross-sectional images from ultrasound B-scans. The parameters that can be directly measured are length, arc length, perimeter, and area. Other parameters such as volume and indicators of cardiac function can be computed from the measured variables.

A block diagram of the system is shown in Fig. 5. The video input image to be processed is usually derived from either a video disk recorder or the stop-motion mode of a video tape recorder, but it may also be obtained from a camera or other continuous video source. The light pen enables the operator to obtain the x-y coordinates of a series of points in outlining the object to be measured. The light-pen electronics provide the x-y coordinate information to an 8-bit microprocessor which performs both logical and arithmetic operations and outputs the data to the memory controller. The CRT monitor shows the input video image mixed with graphical and alphanumeric data generated by the microprocessor system. An example of the measurement technique using the light pen is illustrated in Fig. 6. Here an LAO (left anterior oblique) equivalent orientation of the left ventricle of a normal individual is displayed. In this picture the apex of the left

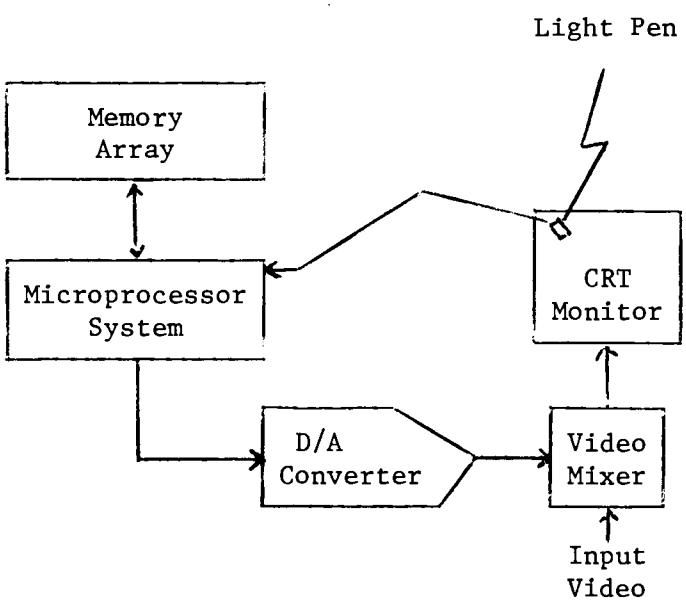


Fig. 5. Block Diagram of Light-Pen System.

Fig. 6. LAO View of Left Ventricle of Normal Patient at End-Systole.

ventricle is at the upper left and the mitral valve appears at the bottom of the dashed line with the left atrium below. By means of the light pen, the left atrium has been encircled and the long axis also indicated. Through the use of a calibration procedure which used calibration marks put on during the original scanning process, the microprocessor system has determined the length of the line and the area enclosed within the outline. The microprocessor has also calculated a volume of 47.9 cc utilizing a standard area-length algorithm.[18] All of the calculated parameters appear on the display along with outlines made with the light pen and the original video B-mode scan. This image occurred at end-systole as selected by the ECG trace below the cardiac image. The point selected in this case was the downward slope of the T-wave. Figure 7 was made later in the same cardiac cycle at the QRS complex and represents the heart at end-diastole. Again the left ventricle was outlined with the aid of the light pen and the long axis indicated. Using the same algorithm, the microprocessor calculated the end-diastolic volume to be 132.9 cc. From these numbers one can readily calculate a stroke volume of 85 cc and an ejection fraction of 0.64. It is apparent from the fact that the midline has shifted from the end-systolic to end-diastolic picture that one cannot use the edges of the outline to determine segmental wall motion. The same problem of heart movement also occurs in quantitative angiography and several methods have been proposed to remove its effects.[19] One method is to superimpose the mid positions of the long axis chords. This has been done in Fig. 8. Further experimentation will be required to determine whether this technique will be useful for the study of segmental wall motion by ultrasound.

Fig. 7. LAO View of Left Ventricle of Normal Patient at End-Diastole.

The same process illustrated above for a healthy individual is next illustrated for a person who has had a myocardial infarction. The infarction is located at the inferior diaphragmatic wall, appearing on the left side of the outline shown in Fig. 9. The cardiac function is borderline with a stroke volume of 30.3 cc and an ejection fraction of 0.44. One can note from the two outlines the region of minimal wall motion is in the region of the infarct.

Although these results are encouraging,[20] one should not draw the conclusion that this technique has

Fig. 8. End-Systole and End-Diastole Views Superimposed.

Fig. 9. End-Systole and End-Diastole Views of Patient with Myocardial Infarct.

been proven. A full clinical evaluation will be required to develop the best procedure and determine the reliability of the information.

Tissue Characterization

The studies reported in this section were undertaken by Dr. William Beaver at Stanford University and a preliminary report was given at the Second International Symposium on Ultrasonic Tissue Characterization.[21] The data were generated using a Varian V-3000 Ultrasonograph which has been modified to permit the rf signal to be coupled to an external system for further processing. This output signal is digitized at the 20-megabit-per-second sample rate and stored in a Biomation 8100 transient recorder. The data gathering system is controlled by a minicomputer system and the data are fed onto a 9-track magnetic tape recorder for further analysis on a large computer. Figure 10 is a block diagram of the system. The ultrasonograph permits one to obtain a real-time image of a desired region and then through the minicomputer control, one fan-beam line from the picture can be digitized and stored for further analysis. A photograph of the B-mode picture permits one to correlate the data from the selected scan line with the actual surroundings. One goal of this study is to analyze features in the radio-frequency signal that do not show up on the B-mode display. Suitable processing may highlight subtle features. Another goal is to study characteristics of signals from known healthy and diseased tissues *in vivo* to attempt to determine tissue signatures.

Before looking at tissue samples, it is a good idea to investigate the characteristics of the data gathering system. The impulse response of the system has been determined by reflecting a signal from a wire target. The digitized signal has been stored and Fourier analyzed off line. The results are illustrated in Fig. 11. The curve on the left-hand side is the rf signal as it emerges from the ultrasonograph. The Fourier transform is displayed as an amplitude vs. frequency curve on the right of Fig. 11. The bandwidth as determined by the half-amplitude points is approximately 1.1 MHz permitting one to obtain a resolution of approximately 0.7 mm in depth.

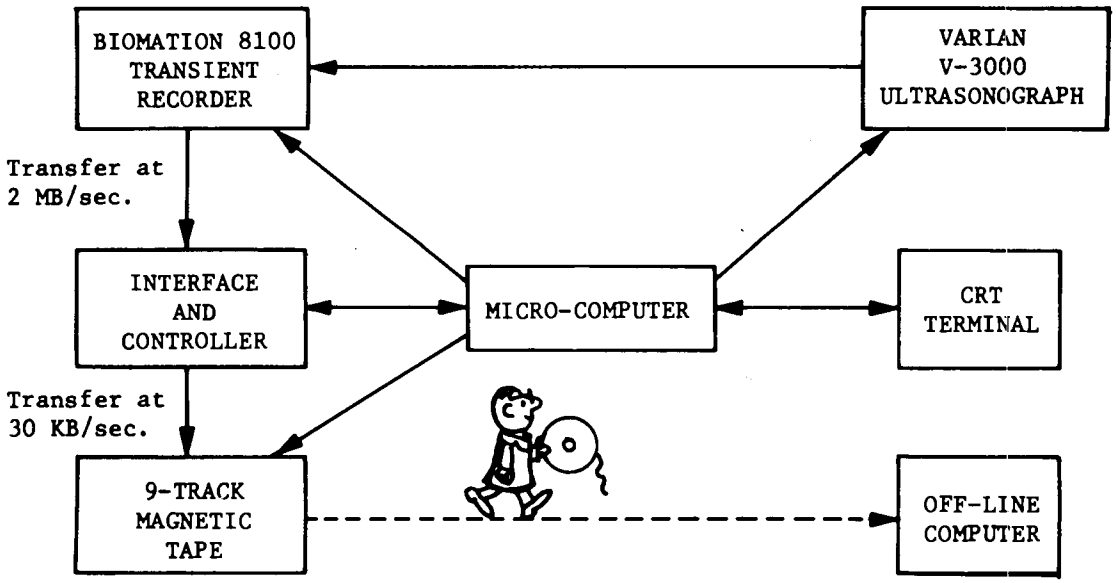


Fig. 10. Block Diagram of Computer-Controlled System

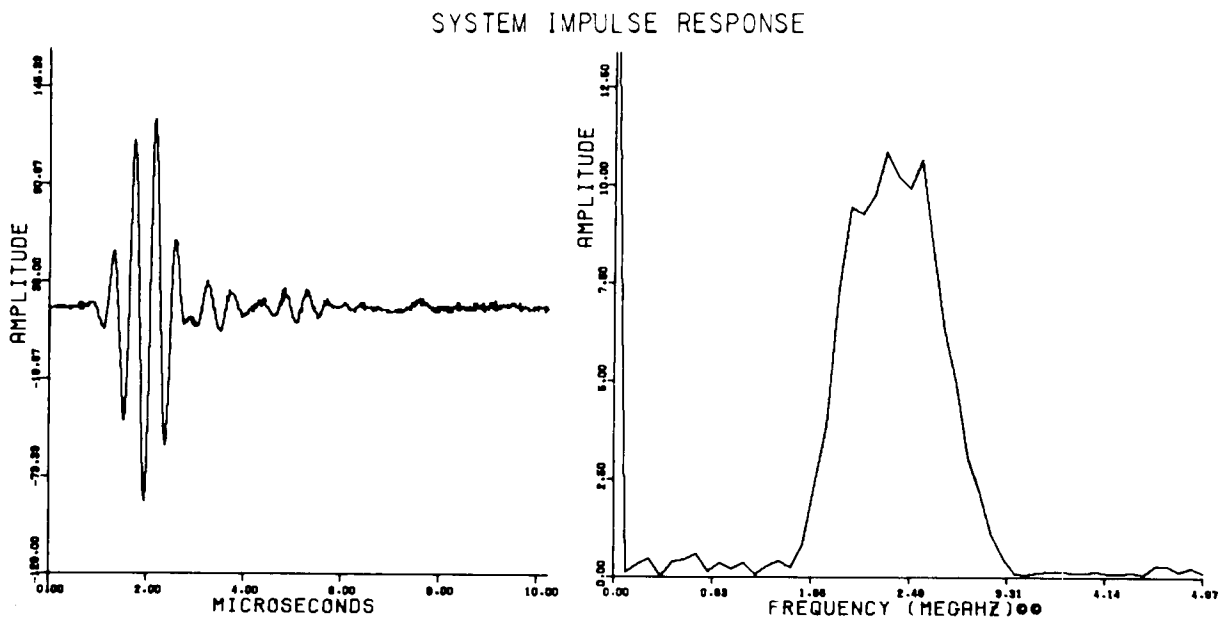


Fig. 11. System Impulse Response

The application of the technique to the examination of heart tissue is illustrated in Fig. 12. Here the cardiac septa of two normal individuals are studied and the spectra and their Fourier transforms are obtained. It is interesting to note the rather large variations of the Fourier spectra of these two individuals which are indicative of the difficulty of the problem being approached. Only that portion of the rf signal that corresponds to that obtained from the cardiac septa are displayed and transformed.

Not only is there a variation between individuals, but variations also occur with time during the cardiac cycle in the same individual. Figure 13 illustrates a series of four samples taken at the same position but at two-millisecond time intervals. One notes marked changes in the Fourier patterns in these relatively

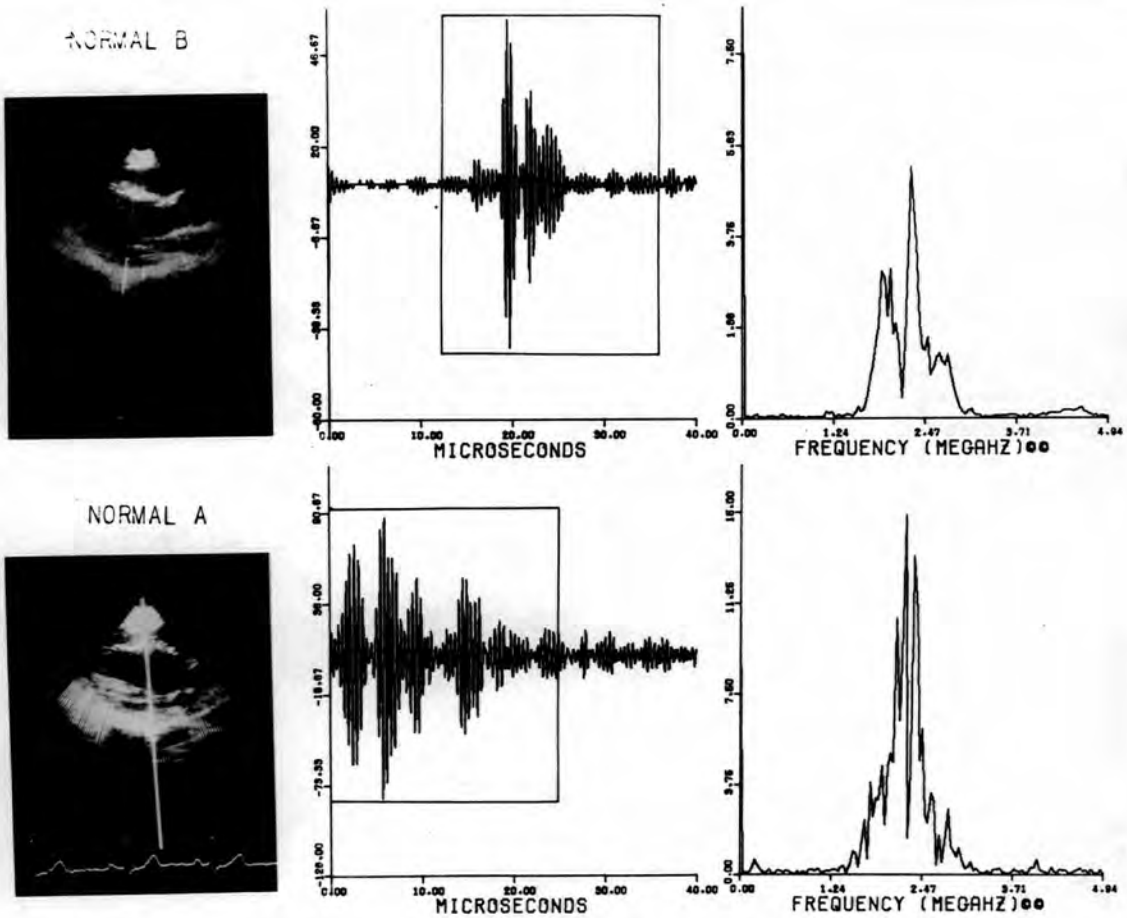


Fig. 12. Spectra and Fourier Transforms of Two Normal Patients

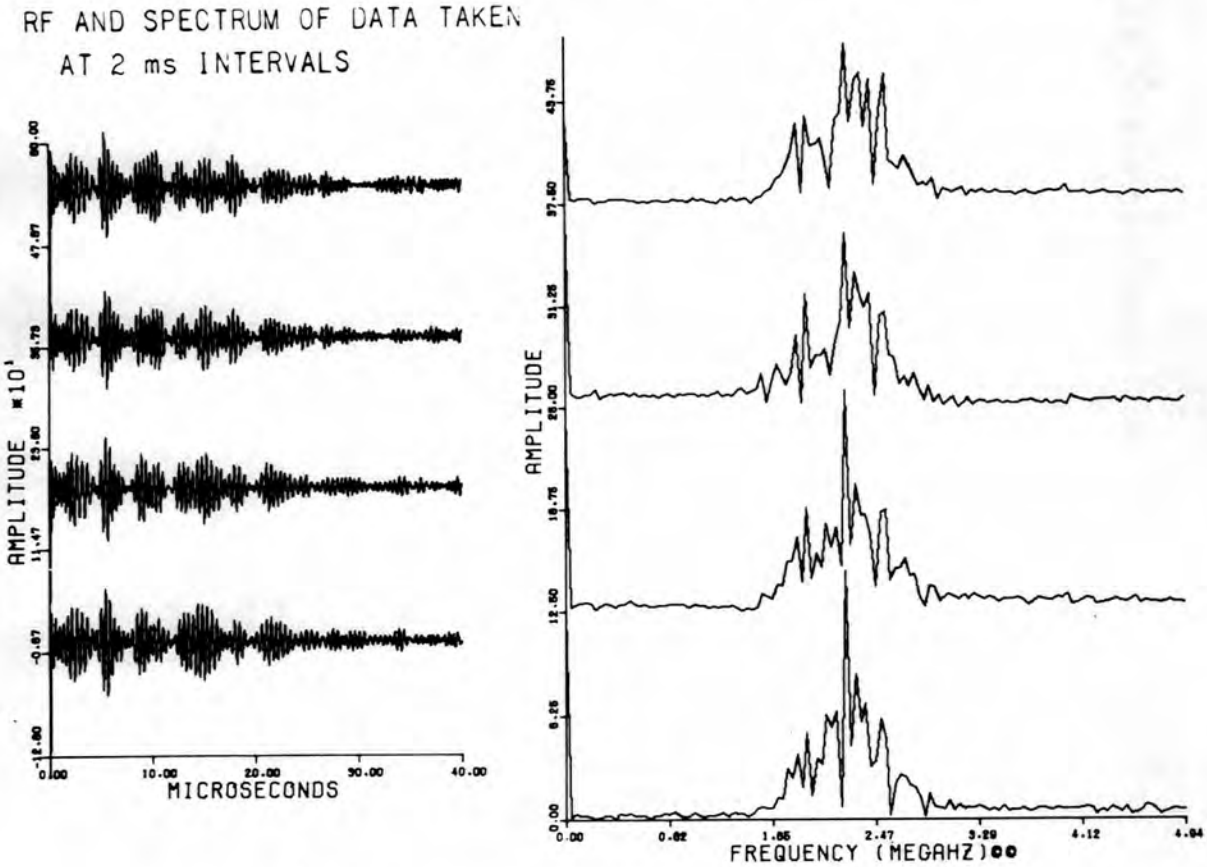


Fig. 13. RF and Spectrum of Data Taken at 2 ms Intervals

short time intervals. Finally, in Fig. 14 four different Fourier components are plotted as a function of time. On the left where the interval in between reference samples is two milliseconds, and on the right the interval is 30 milliseconds. The total time represented in the right-hand side of the figure is approximately one cardiac cycle.

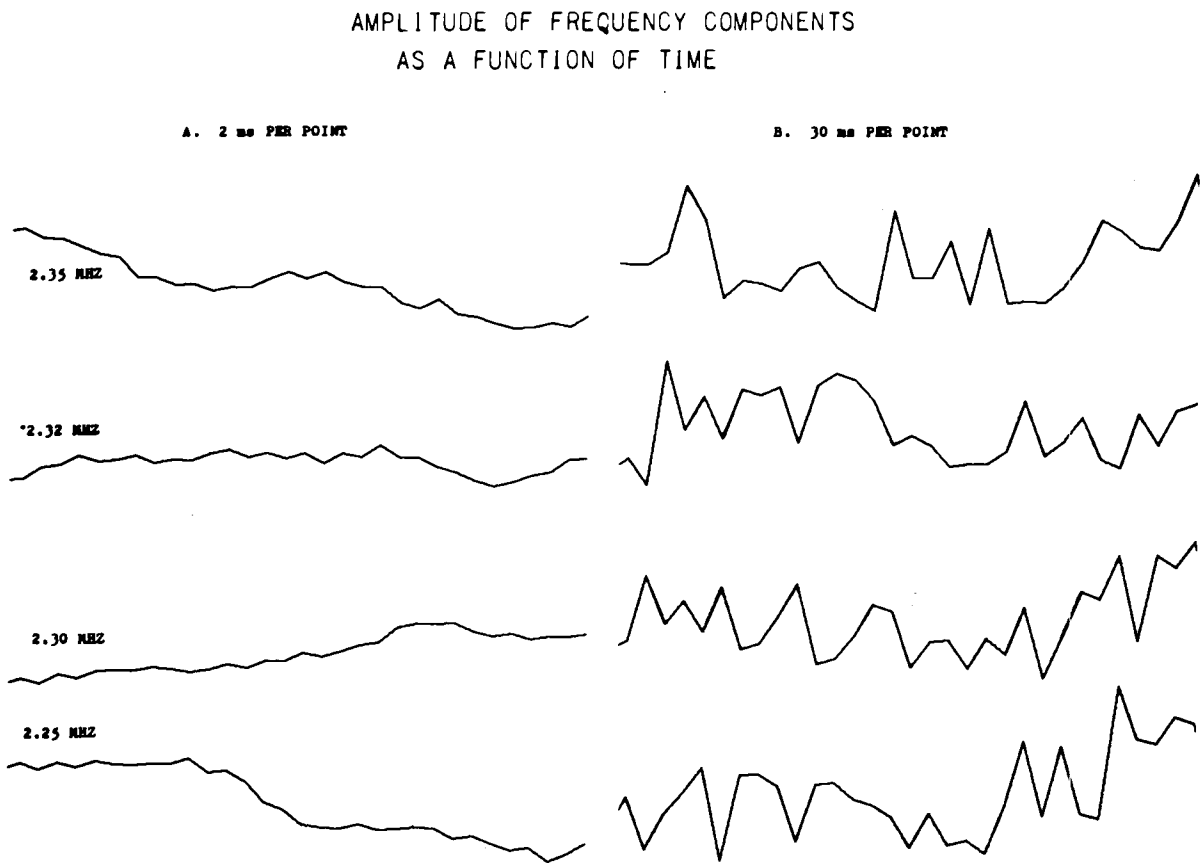


Fig. 14. Amplitude of Frequency Components as a Function of Time.

The initial results indicate that rather subtle changes in tissues can be detected. Further work will be required to develop optimum signal processing techniques and to relate the signal characteristics to normal and pathological tissues.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the work of Dr. Ron Daigle and Mr. William Painter of Varian in the development of the light-pen system; Dr. Nelson Schiller of the University of California, San Francisco, for the data used in Figs. 6 through 9; and to Dr. William Beaver of the Stanford University Integrated Circuits Laboratory for Figs. 10 through 14.

REFERENCES

1. F. A. Firestone, "The Supersonic Reflectoscope, an Instrument for Inspecting the Interior of Solid Parts by Means of Sound Waves," J. Acoust. Soc. Amer. 17, 287 (1945).

2. I. Edler and C. H. Hertz, "Use of Ultrasonic Reflectoscope for Continuous Recording of Movements of Heart Walls," Kurgl Fysiogr. Sallad i Lund Forhandl., 24, 5 (1954).

   I. Edler, A. Gustafson, T. Karlesfors, B. Christensson, "Ultrasound Cardiography," Acta Med. Scand. 170 (Suppl. 370), 67 (1961).

3. J. Griffith and W. L. Henry, "A Sector Scanner for Real Time Two-Dimensional Echocardiography," Circulation, XLIX, 1147 (June 1974).

4. H. H. Holm, J. K. Kristensen, J. F. Pedersen, S. Hancke, A. Northeved, F. Jensen, "A New Mechanical Real Time Ultrasonic Scanner," Ultrasound in Medicine, 2, 459 (1976).

5. J. J. Wild, J. M. Reid, "Application of Echo-Ranging Techniques to the Determination of Structure of Biological Tissues," Science 115, 226, (Feb. 29, 1952).

6. M. G. Maginness, J. D. Plummer, J. D. Meindl, "An Acoustic Image Sensor Using a Transmit-Receive Array," Acoustical Holography, 5, 619 (1974).

7. N. Bom, C. Lancee, J. Honkoop and D. Hugenholtz, "Ultrasonic Viewer for Cross Sectional Analysis of Moving Cardiac Structures," Biomedical Engineering, 500, (Nov. 1971).

8. J. C. Somer, "Electronic Sector Scanning for Ultrasonic Diagnosis," Ultrasonics, 153, (July 1968).

9. N. B. Schiller and N. H. Silverman, "Apex Echocardiography: A New Method for Imaging the Adult Heart using an 80° Phased Array Real-Time Two-Dimensional Sector Scanner," Amer. J. Cardiology, 39 #2, 266, 26th Annual Scientific Session of American College of Cardiology, March 7-10, 1977, Las Vegas.

10. B. W. Gilbert, R. A. Schatz, O. T. vonRamm, V. S. Behard, J. A. Kisslo, "Two-Dimensional Echocardiographic Definition of Mitral Valve Prolapse," Ultrasound in Medicine, 3A, 87 (1977).

    B. W. Gilbert, R. A. Schatz, O. T. vonRamm, V. S. Behard, J. A. Kisslo, "Mitral Valve Prolapse, Two-Dimensional Echocardiographic and Angiographic Correlation," Circulation, 54, No. 5 (Nov. 1976).

11. N. H. Silverman and N. B. Schiller, "Apex Echocardiography: A New Method for Detecting Congenital Heart Disease using an 80° Phased Array Real-Time Two-Dimensional Sector Scanner," Amer. J. Cardiology, 39 #2, 262, 26th Annual Scientific Session of Amer. College of Cardiology, March 7-10, 1977, Las Vegas.

    N. H. Silverman, M. Payot, "Echocardiography in Transposition of the Great Arteries after Mustard's Procedure," Abstract #0175, Circulation, 53 and 54, Suppl. II (Oct. 1976).

12. R. P. Martin, J. W. French, M. M. Pittman, R. L. Popp, "Identification and Localization of Pericardial Effusion by Real Time Two-Dimensional Echocardiography," Ultrasound in Medicine, 3A, 211 (1977).

R. L. Popp, R. P. Martin, J. W. French, M. M. Pittman, "Positional Distribution of Pericardial Effusion by Real Time Two-Dimensional Echocardiography," Abstract #0918, Circulation, 53 and 54, Suppl. II (Oct. 1976).

13. R. P. Martin, J. W. French, M. M. Pittman, R. L. Popp, "Analysis of Idiopathic Hypertrophic Subaortic Stenosis by Wide Angle Two-Dimensional Phased Array Echocardiography," Ultrasound in Medicine, 3A, 249 (1977).

14. J. W. French, R. P. Martin, M. M. Pittman, R. L. Popp, "Ultrasonic Examination of Operative Patients with Conotruncal Abnormalities using a New Wide-Angle Sector Scanner," Ibid, 261.

J. W. French, N. H. Silverman, R. P. Martin, N. Schiller, R. L. Popp, "Examination of Operative Patients with Conotruncal Abnormalities using an Ultrasonic Wide-Angle Scanner," Abstract #0177, Circulation, 53 and 54, Suppl. II (Oct. 1976).

15. J. C. Birnholz, "Screening Studies with Ultrasound," Ultrasound in Medicine, 3A, 747 (1977).

16. J. C. Birnholz, "Routine Pelvic Examinations by Ultrasound," In preparation.

17. J. Stephens, J. C. Birnholz, "The Clinical Significance of Fetal Movement Patterns," Ultrasound in Medicine, 3A, 771 (1977).

J. D. Stephens and J. C. Birnholz, "Modern Management of Fetal Death in Utero and Missed Abortion," Asian Federation of Ob. & Gyn. Proceedings of the 1st Inter-Congress, Singapore, April 27-30, 1976. p. 253.

18. H. T. Dodge, H. Sandler, W. A. Baxley, R. R. Hawley, "Usefulness and Limitations of Radiographic Methods for Determining Left Ventricular Volume," Amer. J. Cardiology, 18, 10 (July 1966).

19. E. L. Alderman, "Clinical Application of a Light-Pen Computer System for Quantitative Angiography," Cardiovascular Imaging and Image Processing, Theory and Practice - 1975, 209, Stanford University SPIE Conference (July 1975).

20. N. Schiller, D. Drew, H. Acquatella, R. Boswell, E. Botvinick, B. Greenberg, E. Carlsson, "Noninvasive Biplane Quantitation of Left Ventricular Volume and Ejection Fraction with a Real-Time Two-Dimensional Echocardiography System," Abstract #0920, Circulation, 53 and 54, Suppl. II (Oct. 1976).

21. L. Joynt, D. Boyle, H. Rakowski and W. Beaver, "Identification of Tissue Parameters by Digital Processing of Real-Time Ultrasonic Clinical Data," 2nd International Symposium on Ultrasonic Tissue Characterization, June 13-17, 1977, Gaithersburg, MD.

DISCUSSION:

D. Ochs indicated that he was always amazed at how much more resolution there seemed to be in the motion pictures compared to the stills. Agreeing, W. Anderson pointed out that it was not better resolution so much as an improved signal to noise ratio in the movie film. He attributed this to two factors: (1) integration by the eye and mind, which are very effective at filtering, and (2) integration by the CRT screen and vidicon during the recording of the images for the movie film.

Following a question by D. Ochs about the resolution of the phased and linear array transducers, W. Anderson indicated that the intrinsic resolution is determined by the aperture of the transducer, and that it has been shown that measured values agree quite closely with the values predicted theoretically. He added that in a phased array the aperture includes all of the elements of the transducer, whereas in the linear array the aperture includes only those elements which are excited at any one time.

An unidentified member of the audience asked what percentage of the adult population in which a study was performed was a diagnostic view of the heart obtained. In reply, W. Anderson stated that his studies were no different in this respect from the standard T-M mode exam, and he believed the figure to be between 75 and 80% of the population.

In response to a question by A. Todd-Pokropek about the method of data collection, W. Anderson replied that he employed a sampling rate of 20 mHz to digitize the echoes along one of the M-lines for a period of approximately 2 millisecs.

B. Pullan queried that, in view of the fact that Fourier transforms were notoriously sensitive to noise, whether the changes in the Fourier transforms in figure 13 could not be accounted for simply by noise in the system. W. Anderson replied that Dr. W. Beaver had made the Fourier transform measurements and felt sure that noise was not a problem.

QUANTITATIVE IMAGING FROM TRANSMISSION ULTRASOUND

J. F. Greenleaf, S. A. Johnson and B. Rajagopalan
Department of Physiology and Biophysic
Mayo Foundation
Rochester, Minnesota, U.S.A.

ABSTRACT

Ultrasonic computer-assisted tomography (UCAT) provides the capability of quantitatively measuring the two-dimensional distributions of various acoustic parameters *in vivo* or *in vitro* under certain conditions. Quantitative distributions of temperature, acoustic speed, fluid velocity, fluid mixing, and qualitative distributions of attenuation have been obtained in isolated tissues. Temperature coefficients of acoustical speed within several tissues have been obtained using reconstruction techniques. Applications include the determination of distributions of acoustic parameters within the breast for diagnosis of breast cancer. Preliminary results in isolated breast tissues are reported which indicate that malignant lesions can be detected utilizing UCAT to measure regional distributions of acoustic speed within the breast. Synergistic relationships between various imaging modalities are described.

Historically, ultrasound has been used to obtain echo-mode images such as B-scans, time motion (TM) scans such as echo-cardiograms and transmission images such as C-scans, all of which correspond to *qualitative* mapping of tissue interfaces and geometries. Images representing accurate *quantitative* distributions of acoustic properties of tissues such as those recently obtained in this and other laboratories[1-5] should be intrinsically more valuable than qualitative images since fundamental tissue properties can be calculated from such data. Quantitative images which represent two-dimensional distributions of mechanical tissue properties can be obtained with acoustic computer assisted tomography techniques.[1,2] The purpose of this paper is to describe methods of obtaining quantitative images of several tissue properties through the use of acoustic computer assisted tomography and to relate the use of these images to both laboratory investigations and clinical applications.

The qualitative images such as the TM image obtained in echocardiography, the B-scan - both compound and linear - obtained with present-day clinical equipment and transmission C-scan images obtained with the transmission camera developed by Green et al.[6],

are methods of obtaining images which are very complex functions of scatter, acoustic impedence, reflection coefficients and other properties of tissue-ultrasound interaction.

Ultrasonic computer assisted tomography is a technique of analysing acoustic data to calculate quantitative two-dimensional distributions of values of acoustic parameters such as acoustic speed and attenuation which can be related to compressability and density or to absorption and scatter respectively.

The equations and geometry necessary for time-of-flight (TOF) reconstruction which results in values of acoustic speed within a two-dimensional plane are shown in Figure 1. The equation in

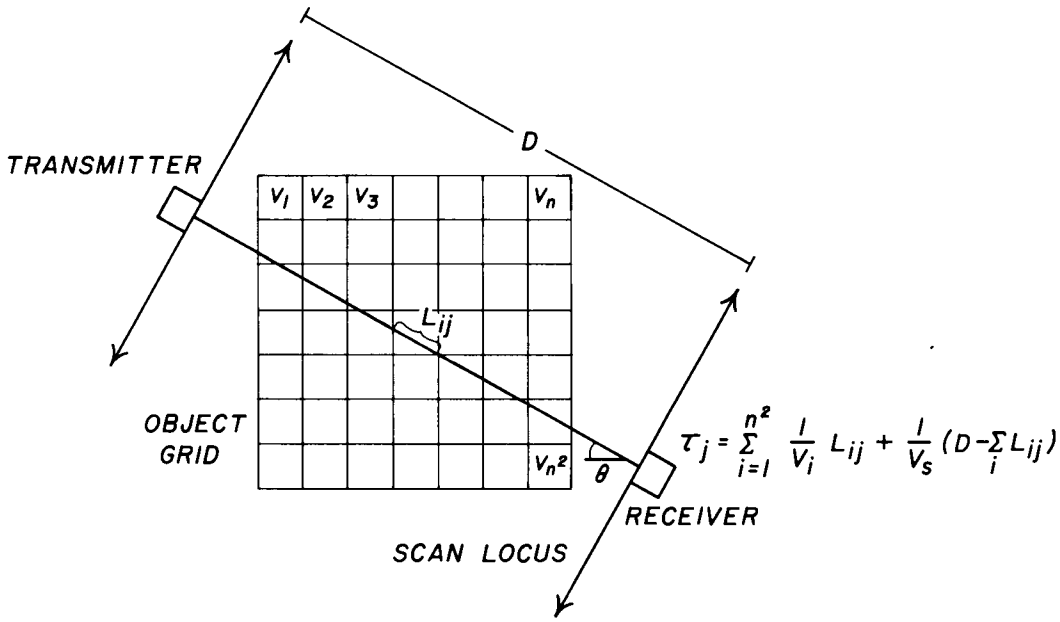


Fig. 1. Geometry of ultrasound transmission for algebraic reconstruction. Grid is fixed to coordinates of object. Speed of sound V_i, is assumed to be constant and isotropic within each pixel. L_{ij} can be varied according to curvature of ray path, which need not be straight. Equation represents propagation delay of pulse through reconstruction region (grid). Number of measurements (i.e., equations) required for solution of set of linear equations can be obtained by scanning object at several angles of view.

Figure 1 represents the relationship of acoustic speeds throughout the tissue to measured transit time of the acoustic pulse, and can be generalized as shown in equation 1 to indicate that any scaler parameter Q which is related through a line intergral to measureable values $P(\theta,\alpha)$ in the received acoustic pulse can be reconstructed using any of several reconstruction techniques (7).

$$P(\theta,\alpha) = \int Q \, dl. \qquad (1)$$

Most of the results shown in this paper are acquired using fan-beam geometry similar to that utilized in several medical x-ray scanners and in our laboratory is accomplished with two transducers, one on either side of the area being examined, a transmitter on one side and a receiver on the other side, moving in the geometry shown in Figure 2. Figure 3 is a block diagram of the laboratory system which enables the computer to control the mechanical and electronic acquisition elements of the scanning equipment. The time-of-flight clock measures arrival times of the acoustic pulses to within 10 nsec and generates a 16-bit digital word representing the arrival time of each pulse. The data are utalized by a fan-beam reconstruction program (8) to obtain quantitative two-dimensional dimensional images of the selected acoustic parameter.

Figure 4 indicates the accuracy of the technique in obtaining distributions of acoustic speed within a test object made of finger cots filled with saline of various concentrations. It is clear from this figure that accuracy on the order of $\pm$ 0.2% can be obtained using time-of-flight techniques for reconstructing acoustic speed.

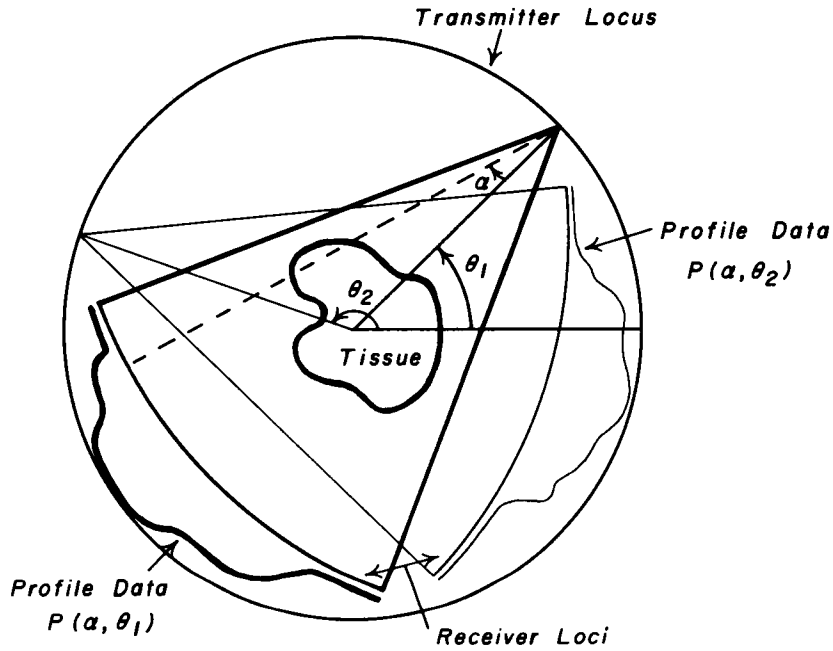


Fig. 2. Fan-beam geometry used for data collection. Transmitter locus and receiver loci are shown for two positions of transmitter. Data are obtained as a function of projection angle θ and beam position α. Theta is stepped in 3° increments over 360° and α is stepped in 0.15° increments over 300 to 400 steps for each view.

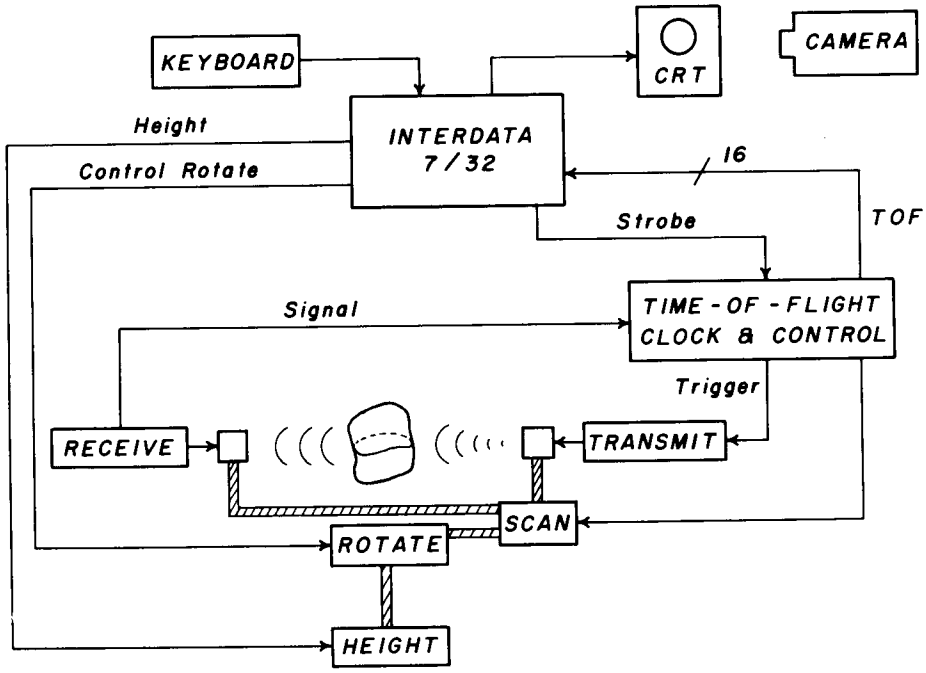
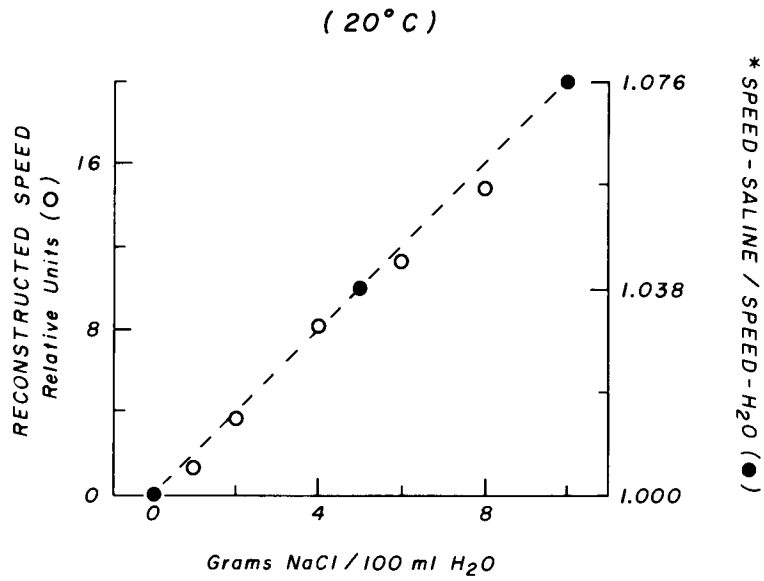


Fig. 3. Schematic of computer-controlled data acquisition system which obtains digitized ultrasonic arrival times from transmission scans to be used for reconstruction. Pulse is triggered by computer at 200-400 points along the fan-beam scan length of about 60°. After analog determination of arrival time or amplitude parameters, data are input to computer for each projection. The scan is then rotated 3° by stepper motor and the process is repeated through 360°. Data are then used by reconstruction algorithms to calculate quantitative two-dimensional distributions of various parameters such as acoustic speed, attenuation, temperature or fluid velocity.



*Millero and Kubinski, J Acoust Soc Am 57(2): 312-319, 1975

Fig. 4. Reconstructed values of speed of sound within rubber finger cot test object containing different saline concentrations compared to measured values. Five values in the center of each finger cot in reconstruction image were averaged to obtain values shown by circles in this graph. Filled circles indicate acoustic velocities for selected saline concentrations obtained from literature. Precision of results are due to ability to measure pulse arrival time to 1 part in 10^4, allowing extremely precise measurements of refraction index (right ordinate).

An example of a reconstruction of the distribution of acoustic speed within an isolated organ - the canine heart - is shown in Figure 5. The apparent resolution of such reconstructions is on the order of 2-3 mm. Although such reconstructions cannot be obtained in the intact animal, the reconstruction of canine hearts is used to illustrate the geometric accuracy of the method.

Fig. 5. Reconstructions of speed of sound within canine heart (left) compared to photograph of section through corresponding level (right). Reconstruction image of 64 X 64 array of pixels was lineraly interpolated up to 128 X 128 pixels. Papillary, epicardial and endocardial surfaces are represented with good fidelity; however, detail shown here of acoustic speed within the myocardial tissue is not yet substantiated.

The distribution of attenuation, calculated from changes in the area of the first wave of the received pulse, is shown in Figure 6 for an excised human female breast scanned in the coronal plane.

Since acoustic speed is affected by temperature, one can use acoustic speed reconstruction techniques to measure distributions of temperature within objects of homogenous acoustic speed.[8,9] Differential changes in temperature can be related to changes in acoustic speed in inhomogenous tissue.

Figure 7 represents distributions of acoustic speed in a homogeneous material, water, containing regions of different temperature. Rubber finger cots were perfused with water of temperatures differing by a few degrees from the ambient water temperature. Most tissues have acoustic speed temperature coefficients near that of water, indicating that equivalent temperature measurement sensitivities could be obtained in tissues.

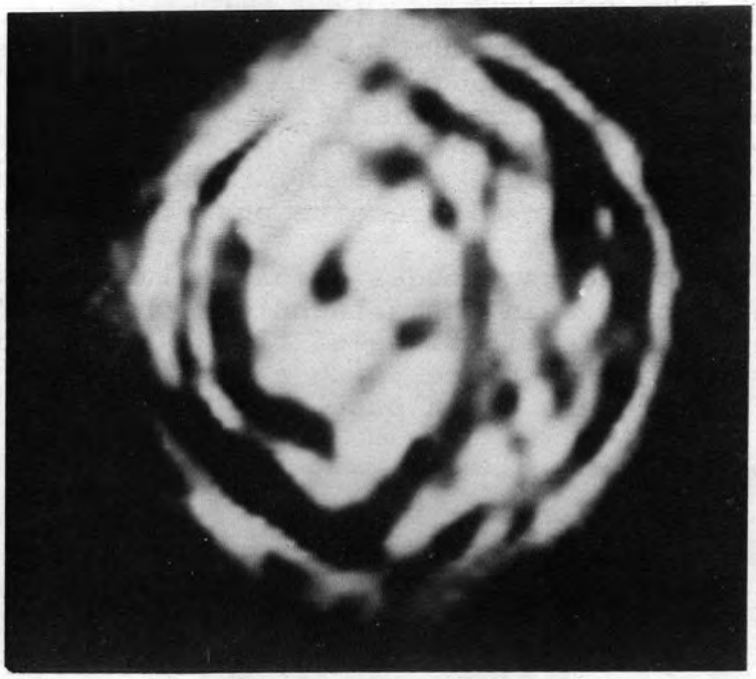


Fig. 6. Distribution of acoustic attenuation in an excised human female breast scanned at 22° C. "Attenuation" was calculated as the area of the first wave received by the transducer. The image is probably not as quantitative as others because attenuation is not rigorously related to area of the pulse.

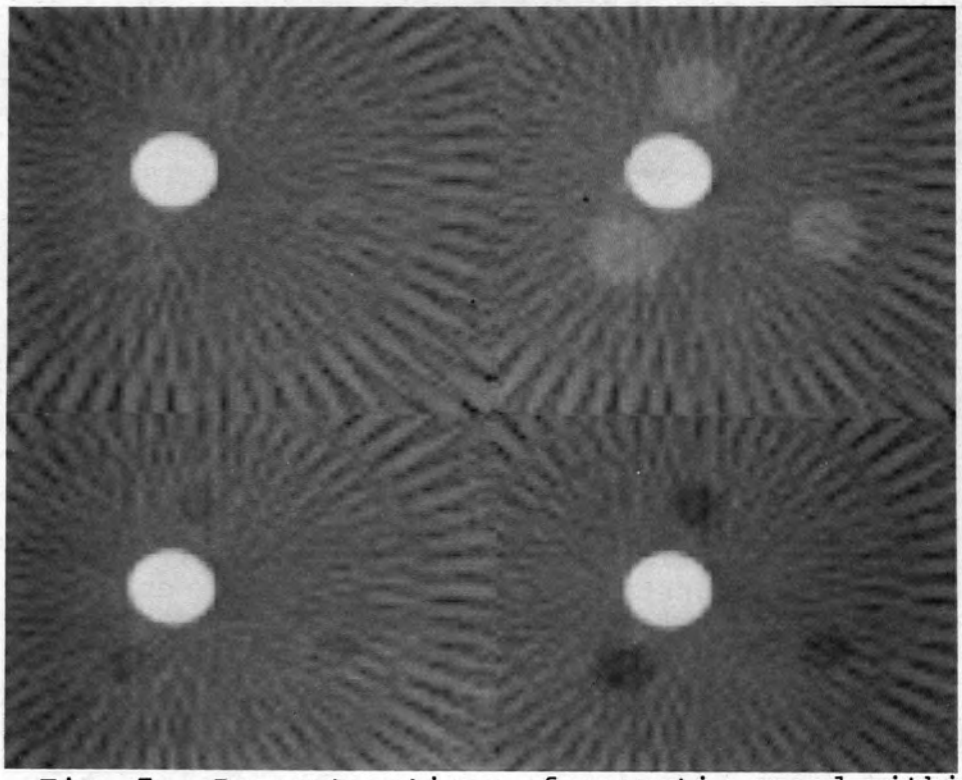


Fig. 7. Reconstructions of acoustic speed within water-filled tank containing one finger cot filled with 2.5% NaCl (bright circle) and three finger cots perfused with water at ambient temperature (T) upper left; T + 1.5° C upper right; T - 0.5° C lower left; and T - 1.0° C lower right panel.

Until recently, vector fields could not be reconstructed but a new technique for reconstructing vector fields has been developed in our laboratory.[8,9] Through the use of time-of-flight measurements, reconstructions of distributions of velocity of fluid within the sampling space can be accomplished using measurements of transit time for the acoustic pulse traveling in both directions through the tissue for each transmit-receive position.

Figure 8 represents the design of a fluid velocity experiment and resulting data. Using this technique, both distributions of acoustic speed and distributions of fluid velocity can be obtained simultaneously within a plane using a single scan without the necessity of scatterers as in Doppler methods.

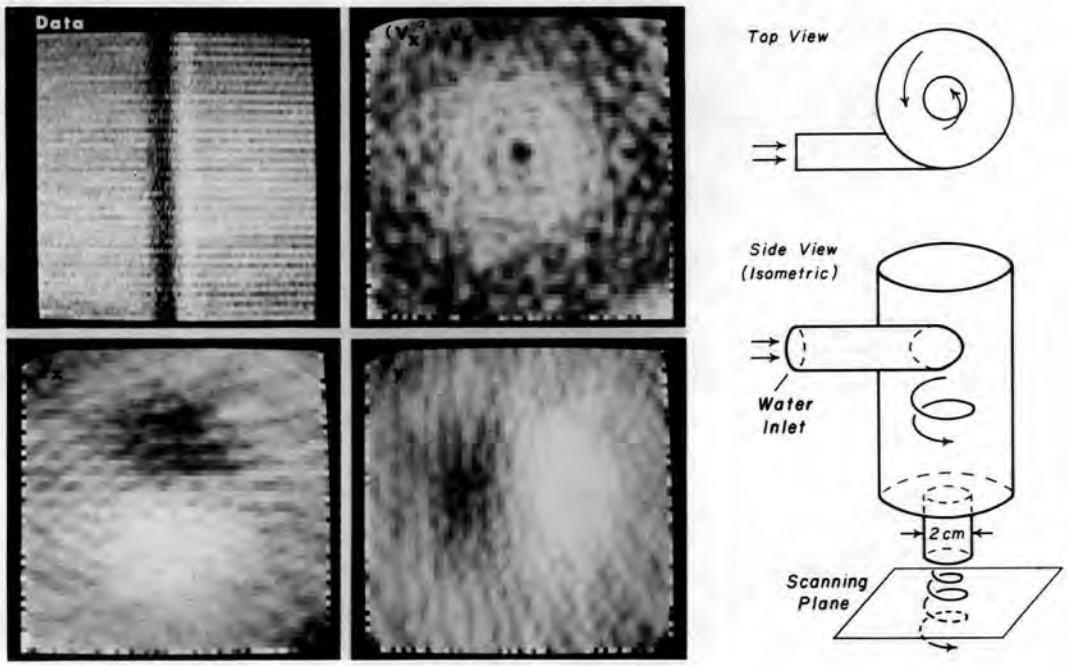


Fig. 8. Experimental data and reconstruction of vector components of fluid vortex. Top left shows experimental data. Black is negative, gray is zero, white is positive. Top right shows reconstructed planar fluid speed $(V_x^2 + V_y^2)^{1/2}$, black is zero, white is positive. Bottom left shows x component of velocity V_x. Bottom right shows y component of velocity V_y. In V_x and V_y, black is negative, gray is zero, and white is positive. Reconstructed flow is maximum (73 cm/sec) at a radius of 0.62 cm. Reconstructions are 64 pixels per side. Geometry of vortex, scanning plane, and vortex generator are shown in top and side views in right margin.

The capability of generating quantitative images using ultrasonic computer assisted tomography can be applied to many areas, both in the experimental laboratory and in the clinical environment. For instance, all compound B-scans are blurred by the inhomogeneity of acoustic refraction throughout the area under investigation. Since acoustic speed varies within the tissue intervening between

the transmitter, the target, and the receiver, depending on the view from which the target is insonified, the arrival time of the resulting echo will vary inversely with the intergal of the acoustic speed along each echo path. Such an effect causes blurring of the resulting compound B scan images.

Figure 9 is a computer simulation of the kind of blur to be expected from a perfect point reflector centered within a human female breast and scanned with a compound B-scan over 2 π radians.

Computer Simulated Compound B-Scan *Reconstruction of Acoustic Speed within Breast*

Fig. 9. Left panel: computer calculated compound B-scan of a perfect reflector centered within the coronal plane of an excised femal breast (16 B-scans equispaced around 360°). Arrival times of echoes received from various angles were calculated by integrating the reciprocal of the reconstructed velocity values along the ray path (from transducer to reflector to transducer) with the reconstruction of acoustic speeds in an excised breast shown in the right panel.

The distribution of acoustic speed (shown in the right panel) was measured from an isolated breast using ultrasound computer assisted tomography, and was used to calculate the arrival times of echoes received from a hypothetical perfect point reflector centered in the breast. The time of arrival of the acoustic pulse received from the point reflector within each B-scan varied depending on the viewing angle of the B-scan transducer. These alterations in arrival times cause a blur of the point reflector which is seen in Figure 9. The distance of the blur is approximately 4 mm since each pixel is 2.2 mm on a side. This indicates that the highest resolution possible at any frequency is approximately 2 or 3 mm in the center of a normal breast if the alterations of arrival times due to inhomogeneous acoustic speeds within the breast are ignored during the acquisition of a compound B-scan. Therefore, an important use of acoustic reconstruction would be in the correction of compound B-scans so that blurring caused by alterations in acoustic speed would be minimized.

Another use for acoustic reconstruction techniques is that of tissue characterization of isolated tissues. Blocks of isolated tissues selected from various organs of the body were inserted into finger cots while emersed in degased saline. The finger cots were tied and emersed in a scanning tank and data for two-dimensional reconstructions of acoustic speed were obtained within transverse planes through groups of tissue samples as shown in Figure 10. By altering the temperature of the emersing bath, one can study the effect of temperature on acoustic speeds of various tissues. Since acoustic reconstruction methods obtain average values for acoustic speed over all angles of view around the tissue, the affects of anisotropicity of acoustic speed on the measurement are minimized.

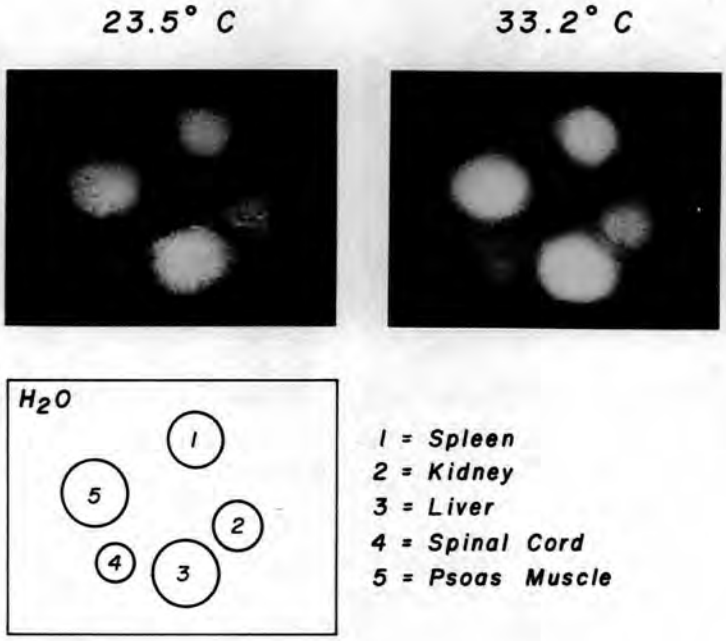


Fig. 10. Reconstruction through tissue samples suspended in scanner water tank. Two different images show effect of temperature on acoustic speed within individual tissues.

Another application of acoustic reconstruction is in the area of breast diagnosis. The availability of a 2 π radian aperture around the breast allows one the utilize acoustic reconstruction to obtain distributions of attenuation and velocity and other parameters within the intact breast. Figure 11 represents distributions of acoustic speed and attenuation in an isolated breast and includes associated optical photographs of the surface of the breast sliced at the level which was scanned by the acoustic scanner. The relationship between the anatomy and the acoustic speed is quite clear and when added to the image representing distribution of acoustic attenuation provides a synergistic depiction of distributions of the two acoustic properties within the tissue.

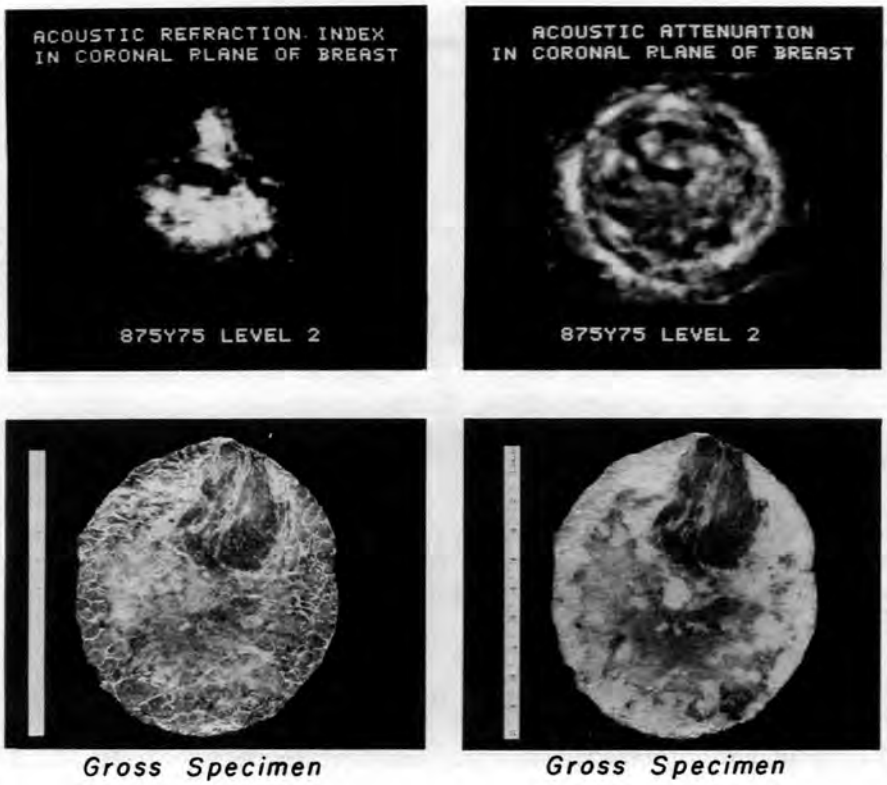


Fig. 11. Reconstruction of speed of sound (upper left) and acoustic attenuation coefficients (upper right) within excised human breast tissue obtained from postmortem subcutaneous mastectomy. Photographs of surface of specimen sectioned at plane of acoustic scan is shown in lower panels. Scan data were obtained in a manner described in text. Both arrival time and amplitude profiles were calculated from the digitized pulses to obtain sound speed and attenuation images, respectively. Although these images are uncorrected for refraction errors, the detail seems promising. Lower region of increased velocity (CA) corresponds to infiltrating medullary carcinoma while upper region (M) corresponds to muscle. Fat (F) appears as low velocity regions. High attenuation is associated with connective tissue regions near muscle.

The use of acoustic reconstruction for obtaining quantitative distributions of tissue parameters in vivo will probably be limited to those areas of the body which are amenable to obtaining a 2 π radian angle of view. Such areas of the body are extremely limited in number. Therefore, acquisition of quantitative data from echo data is highly desired. Such techniques are well known in seismology and other areas and may provide a basis for developing methods of obtaining quantitative images of several tissue properties in areas of the body having limited apertures.

REFERENCES

1. J. F. Greenleaf, S. A. Johnson, S. L. Lee, G. T. Herman, and E. H. Wood, "Algebraic reconstruction of spatial distributions of acoustic absorption within tissue from their two-dimensional acoustic projections," Vol. 5, p. 591-603 in Acoustical Holography, Philip S. Green, Editor, Plenum Press, New York, 1974.

2. J. F. Greenleaf, S. A. Johnson, W. F. Samayoa, and F. A. Duck, "Algebraic reconstruction of spatial distributions of acoustic velocities in tissue from their time-of-flight profiles," Vol. 6, p. 71-90 in Acoustical Holography, Newell Booth, Editor, Plenum Press, 1975.

3. J. P. Jones, "Impediography: A new ultrasonic technique for nondestructive testing and medical diagnosis," p. 214 in Ultrasonic International, 1973, Conference Procedding, (PIC Science and Technology Press, United Kingdom, 1974).

4. A. C. Kak, L. R. Beaumont, and J. Wolfley, Signal Processing in the determination of acoustic impedance profiles, School of Electrical Engineering, Purdue University, TR-EE 75-7, (December, 1974).

5. G. H. Glover and J. C. Sharp, "Reconstruction of ultrasound propagation speed distributions in soft tissue: Time-of-flight tomography," IEEE Transactions on Sonics and Ultrasonics, SU-24 (4),229-234 (1977).

6. P. S. Green, L. F. Schaefer, E. K. Jones, and J. R. Suarex, "A new high-performance ultrasonic camera," p. 493-503 in Acoustical Holography, Plenum Press, New York, 1974.

7. R. Gordon and G. T. Herman, "Three-dimensional reconstruction from projections: A review of algorithms," Vol. 38, p 111-115 in International Reviews of Cytology, 1974.

8. S. A. Johnson, J. F. Greenleaf, W. F. Samayoa, F. A. Duck, and J. D. Sjostrand, Reconstruction of three-dimensional velocity fields and other paramenters by acoustic ray tracing, 1975 Ultrasonic Symposium Proceedings, IEEE Cat. #75, CHP 994-4SU (1975).

9. S. A. Johnson, J. F. Greenleaf, M. Tanaka, G. Flandro, "Reconstructing three-dimensional temperature and fluid velocity vector fields from acoustic transmission measurements," Instrument Society of America Transactions, 1977 (In Press).

ACKNOWLEDGEMENTS

The authors thank Miss Pat Snyder and co-workers for secretarial and graphical assistance.

Dr. R. C. Bahn assisted in acquisition and identification of tissues.

This research was supported in part by grants HT-4-2904, HL-04664, HL-18968, RR-00007, HL-00060 from NIH, USPHS; NCI-CB-64041 and NCI-CB-53857 from NCI.

TISSUE CHARACTERISATION BY AN ANALYSIS OF THE DISTRIBUTION OF ATTENUATION VALUES IN C-T SCANS

B.R. Pullan, R. Fawcett and I. Isherwood
Departments of Medical Biophysics and Diagnostic Radiology
University of Manchester

ABSTRACT

Attenuation values as measured by a C-T scanner are distributed in both magnitude and space. The distribution in magnitude can be described by the mean, standard deviation and skewness of histograms of attenuation values. The distribution in space can be characterised by autocorrelation functions or histograms of gradients. Using these measures of scan structure it has been shown that the spatial distribution of attenuation values can be influenced by tissue and that changes in tissue structure corresponding to disease can be detected.

INTRODUCTION

The two dimensional distribution of linear attenuation coefficient measured and displayed by a C-T scanner can be thought of as describing a three dimensional surface as shown in Figure 1.

In the case of a perfect C-T scanner and an object such as water with a completely uniform distribution of attenuation values, the surface would be expected to be flat with no deviations from a true horizontal plane. This is not found to be the case in practical machines and a distribution or spread of attenuation is found.

When tissue is scanned there will be an additional variation in attenuation values resulting from tissue structure. The total variation, or noise, will therefore depend upon both the scanner and the tissues being scanned.

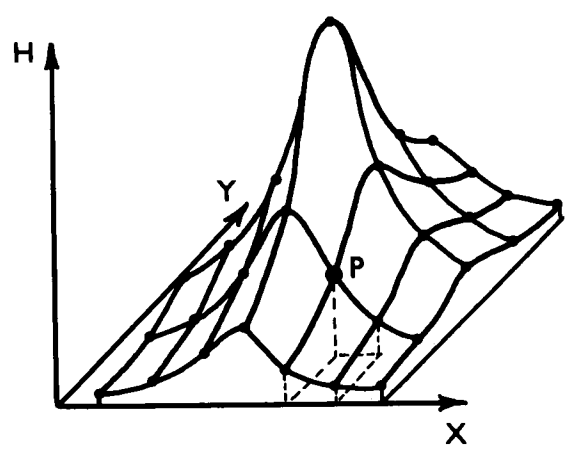


Fig. 1. Representation of attenuation values as a surface.

In this paper methods of characterising noise will be described and it will be shown that significant differences in structure between normal liver and spleen, normal and cirrhotic liver and normal and atrophied brain can be detected by analysing the spatial structure of the noise in regions of C-T scans corresponding to the appropriate tissues.

THEORY

The noise in any region of a C-T scan can be characterised in terms of either its distribution in amplitude or space. The traditional descriptors of amplitude distribution are the moments of the histogram of attenuation values in a region. The traditional descriptors of the distribution of noise in space are the two dimensional auto-correlation function and power spectral density [1]. These latter two are related and it is only necessary to give one of the two functions [2]. In what follows only the auto-correlation function will be used. An additional descriptor of the spatial distribution will be described in which histograms of gradient directions and amplitudes are employed.

Amplitude Distributions

The distribution of attenuation values, for any chosen region of scan, can be described by a histogram. Such histograms can in turn be described, for most practical purposes, by the first three central moments. The first central moment is more widely known as the mean and if the attenuation values in an area of interest are H_1, H_2, H_3.....H_n the mean or first central moment is defined as

$$M1 = \left[\sum H_n\right] / N \qquad (1)$$

The second central moment, more widely known as the variance is defined as

$$M2 = \left[\sum (H_n - M1)^2\right] / N \qquad (2)$$

This is the mean squared deviation from the mean.
The third central moment is defined as

$$M3 = \left[\sum (H_n - M1)^3\right] / N \qquad (3)$$

This is the mean cubed deviation from the mean.

This latter central moment is zero for a symmetrical distribution and increases in magnitude with increasing skewness of the distribution.

It is often more convenient to give values for the square root of the second moment, the standard deviation, and the cube root of the third moment, the skewness.

The distribution of attenuation values in a region does not provide any information about the way in which the values are distributed in space and for this information we have to look to the auto-correlation function or to an

analysis of gradients.

The Auto-Correlation Function

The auto-correlation function is an average measure of the spatial distribution of attenuation values and is equal to the average product of the attenuation values H(X,Y) and H(X,Y) shifted by J in the X direction and K in the Y direction. For the purposes of this paper the auto-correlation function is taken to be

$$R(J,K) = 1/N \sum\sum H(X,Y)H(X+J,Y+K) \tag{4}$$

N is the number of terms contributing to the sum of products and an average product will therefore be produced for each shift J,K. The value of R(J,K) at some particular J and K is a measure of the average relationship between attenuation values at H(X,Y) and H(X+J,Y+K). The value of the auto-correlation function at shift J,K can be thought of as a measure of the conditional probability that given a positive (or negative) value of H(X,Y) at a particular point a value of like sign will appear at shift J,K.

Gradient Analysis

A very simple measure of the spatial structure of noise is provided by an analysis of gradients. The maximum gradient in attenuation values at each point in a scan has both magnitude and direction. The matrix of attenuation values which constitute a C-T scan can therefore be used to derive two matrices, one of maximum gradient directions and one of maximum gradient amplitudes. Histograms of gradient amplitudes and directions are sensitive to changes in the spatial distribution of noise and are used in this paper to describe the spatial character of C-T scanner noise in the presence of different tissues.

The maximum gradient at a point in a scan can be approximated by

$$G(X,Y) = \sqrt{\left[H(X,Y)-H(X+D,Y)\right]^2 + \left[H(X,Y)-H(X,Y+D)\right]^2} \tag{5}$$

and the maximum gradient direction by

$$\theta(X,Y) = \mathrm{Tan}^{-1}\left[\left[H(X,Y)-H(X+D,Y)\right] / \left[H(X,Y)-H(X,Y+D)\right]\right] \tag{6}$$

D is the distance in picture elements, over which the gradient is measured. Histograms of the quantities G(X,Y) can be adequately described by the first three central moments or mean, standard deviation and skewness. Histograms of the quantities $\theta(X,Y)$ can be described in terms of the difference between the histograms which are observed and those expected. The difference is measured by χ^2 where

$$\chi^2 = \frac{\left[E(\theta)\Delta\theta - O(\theta)\Delta\theta\right]^2}{E(\theta)\Delta\theta} \tag{7}$$

$E(\theta)\Delta\theta$ and $O(\theta)\Delta\theta$ are the expected and observed numbers of gradients respectively falling in the ranges of angles θ to $\theta + \Delta\theta$.

The expected number of gradients in each interval is taken to be

$$E(\theta)\Delta\theta = \frac{(N-D)^2 \Delta\theta}{360} \qquad (8)$$

where it is assumed that the matrix being analysed is N elements square and the gradients are measured over a distance D. χ^2 can be calculated for a set of different distance D and a plot of χ^2 against D used to indicate the way in which the anisotropy in gradient direction varies with the distance over which the gradients are measured. In this study $\Delta\theta$ was chosen so that there were 10 values of $E(\theta)$ and $O(\theta)$ between 0 and 180°.

METHOD

Print-outs of C-T scans of four normal patients, four patients with cerebral atrophy and four water scans were obtained from an EMI CT1000 scanner. A small region measuring 16 x 16 pixels was selected on each scan print-out in the position indicated in Figure 2. The positions of the boxes were chosen such that the ventricles and the Sylvian Fissure were always excluded and in each case the boxes were in geometrically similar positions. Print-outs of C-T scans of five patients with normal livers and spleens and five patients with cirrhotic livers were also obtained but from a prototype EMI CT5000 body scanner. Five water scans were also obtained for a 13" diameter water phantom. Small regions measuring 16 x 16 pixels were selected on the print-outs at positions corresponding to liver and spleen tissue as indicated in Figure 3. In each case the data boxes were placed in geometrically similar positions.

Fig. 2. Position of data box in brain.

Fig 3. Position of data boxes in liver and spleen.

Regions were also selected from the print outs of the water scans at geometrically equivalent positions to liver and spleen. In each case the data in the 16 x 16 pixel boxes was typed manually into a computer and stored as files.

Computer programmes were written to process the stored files and determine 1) the mean, standard deviation and skewness of the histograms of attenuation values and gradients of attenuation values measured over different distances, 2) the mean of the auto-correlation function values along the X and Y axes and 3) the values of χ^2 as defined

in equation 7 for gradients measured over different distances.

RESULTS

The Mean, standard deviation and skewness of histograms of attenuation values are given in Table 1 for brain and water, in Table 2 for normal liver, cirrhotic liver, spleen and water. The same measures of distribution of attenuation values are given for histograms of gradients of attenuation values measured over one pixel spacing in Table 3 for brain results and in Table 4 for liver and spleen results.

Table 1. Brain Attenuation Values

Material	Mean	Standard Deviation	Skewness
Water	-1.07 ± 0.03	2.44 ± 0.06	1.14 ± 0.16
Normal Brain	15.67 ± 0.22	3.12 ± 0.06	1.40 ± 0.30
Atrophied Brain	15.67 ± 0.80	3.19 ± 0.22	1.52 ± 0.27

Table 2. Body Attenuation Values

Material	Mean	Standard Deviation	Skewness
Water*	1.90 ± 0.58	9.31 ± 2.40	7.26 ± 1.42
Normal Liver	32.43 ± 2.18	11.08 ± 0.50	3.84 ± 0.35
Cirrhotic Liver	22.98 ± 1.04	10.90 ± 1.02	4.47 ± 0.77
Spleen	25.62 ± 0.89	10.26 ± 0.57	3.69 ± 0.81

* No significant difference between liver and spleen positions

Table 3. Brain Attenuation Gradients

Material	Mean	Standard Deviation	Skewness
Water	3.64 ± 0.10	1.93 ± 0.08	1.62 ± 0.15
Normal Brain	4.25 ± 0.13	2.26 ± 0.09	1.85 ± 0.12
Atrophied Brain	4.18 ± 0.18	2.22 ± 0.07	1.98 ± 0.14

Table 4. Body Attenuation Gradients

Material	Mean	Standard Deviation	Skewness
Water*	16.11 ± 0.45	8.82 ± 0.36	7.88 ± 1.03
Normal Liver	16.62 ± 0.85	9.08 ± 0.60	8.47 ± 0.65
Cirrhotic Liver	17.27 ± 1.83	8.96 ± 0.84	8.38 ± 1.14
Spleen	14.93 ± 0.85	8.19 ± 0.51	7.84 ± 0.59

* No significant difference between liver and spleen positions

No significant differences are observed between the distributions of attenuation values or gradients for normal and atrophied brain. Water does however have significantly lower mean attenuation values and gradients and also lower standard deviations of attenuation values and gradients than both normal and atrophied brain. Significant differences were observed between the mean attenuation values of cirrhotic and normal liver and also between spleen and normal liver although there were overlaps between these groups. The standard deviations and skewnesses were not significantly different. There were no significant differences between the distributions of attenuation gradients for the above groups. No significant differences were demonstrated between distributions of attenuations or gradients for water in positions corresponding to liver and spleen.

Plots of χ^2 as a function of the distance over which gradients were measured are shown in Figure 4 for normal and atrophied brain; the error bars represent the standard error of the mean. The separation of the curves for the two groups of data can be expressed in terms of t

where $t = \frac{\text{difference between the mean}}{\text{probable error of difference}}$

Plots of t as a function of the distance over which gradients were measured are shown in Figure 5 and demonstrate differences which could only have occured by chance less than 1 time in 100 trials.

Plots of χ^2 versus the distance over which gradients are measured are shown in Figure 6 for normal and cirrhotic liver.

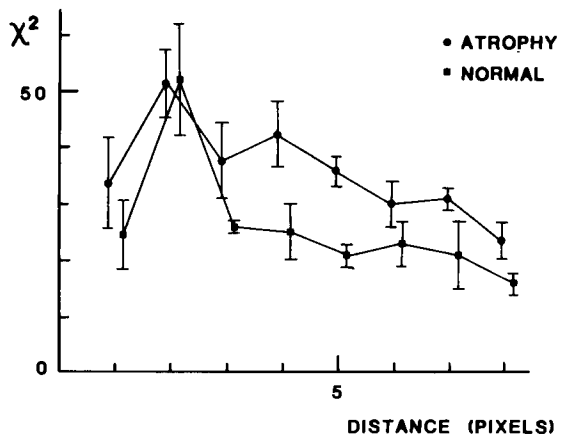


Fig. 4. χ^2 as a function distance.

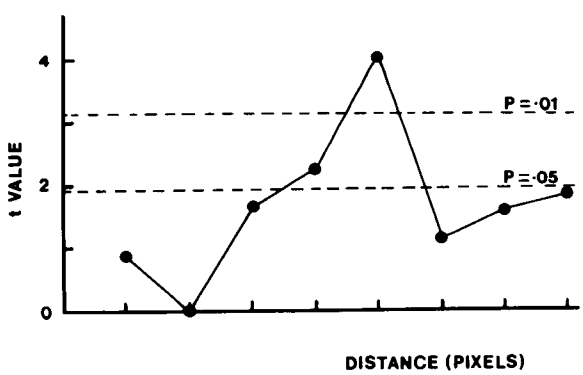


Fig. 5. Separation between normal and atrophied brain in units of standard error

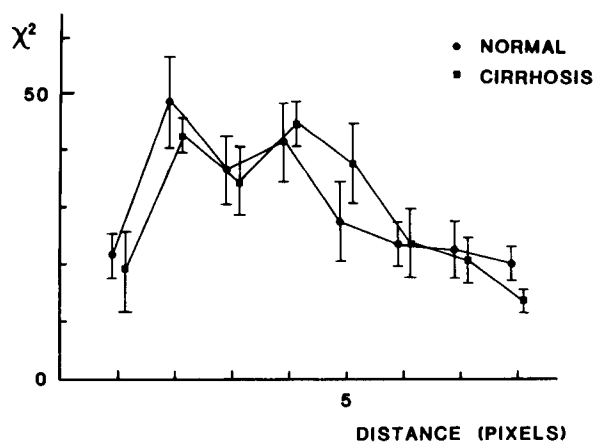


Fig. 6. Separation between normal and cirrhotic liver in units of standard error

Similar curves are obtained for spleen but no significant difference can be detected between the curves for normal liver, cirrhotic liver and spleen.

Autocorrelation functions which have been averaged over two directions at right angles are shown in Figure 7 for both normal and atrophied brain. No significant differences are apparent. Similar auto-correlation functions are shown in Figure 8 for spleen and normal liver and in Figure 9 for cirrhotic and normal liver. The differences between these curves are demonstrated in Figure 10 in which t values are plotted against distance in pixels. Significant differences occur between normal liver and spleen and normal and cirrhotic liver at 3 and 10 pixels spacing respectively. The differences between the auto-correlation functions could only have occured by chance less than one time in 100 trials.

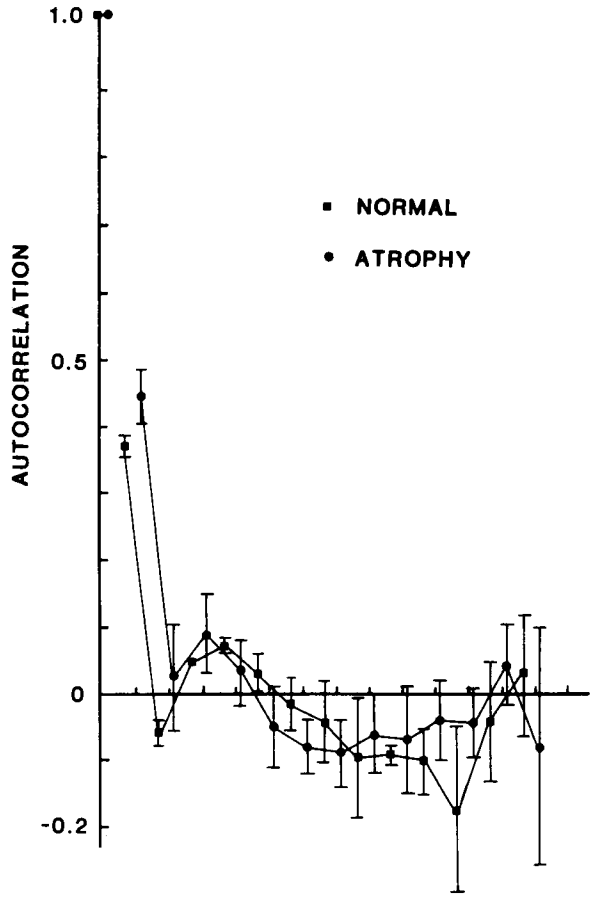


Fig. 7. Averaged auto-correlation functions for atrophied and normal brain.

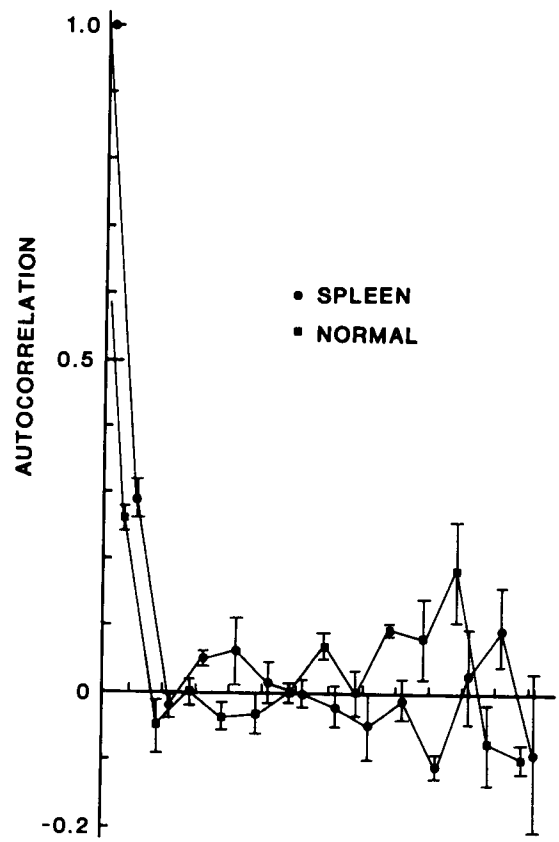


Fig. 8. Averaged auto-correlation function for spleen and normal liver.

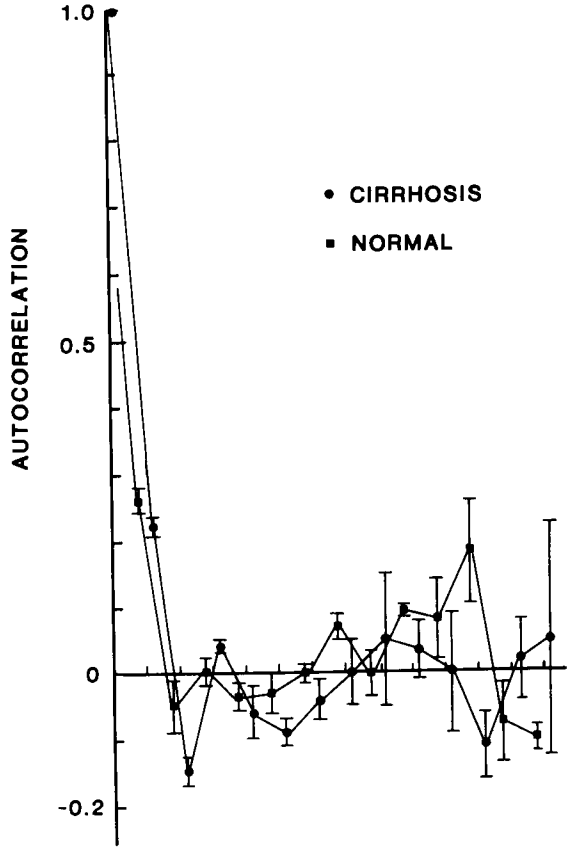


Fig. 9. Averaged auto-correlation functions for normal and cirrhotic liver.

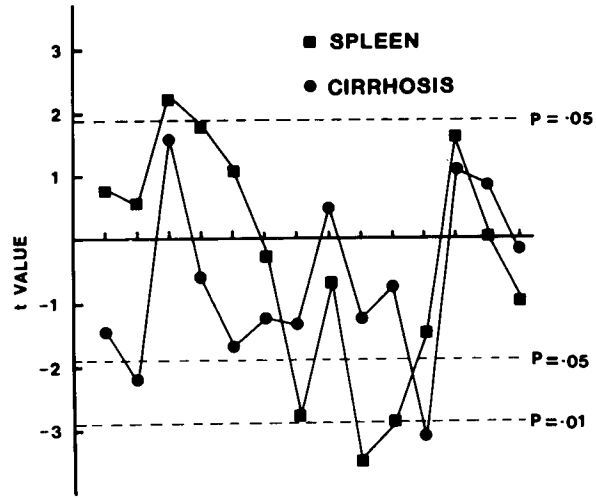


Fig. 10. Separation of spleen and cirrhotic liver from normal liver in units of standard error.

DISCUSSIONS AND CONCLUSIONS

The results indicate that underlying tissue structures in liver, spleen and brain do significantly influence the spatial distribution of attenuation values and that changes in structure which are associated with disease can be detected. In the case of normal and atrophied brain there was a complete separation of the two groups of χ^2 values for gradients measured over a spacing of 5 pixels. There was also a complete separation of auto-correlation function values at ten pixels spacing in the cases of normal liver and spleen and normal liver and cirrhotic liver.

Any conclusions are preliminary in the sense that only a limited number of cases have been included in each group and relatively small matrices of attenuation values have been analysed. Statistically significant differences between different tissues and between normal and diseased tissues can however be demonstrated. It is felt that there is good evidence to support the suggestion that tissue structures and changes in tissue structures can be detected and measured in C-T scans.

REFERENCES

1. P.A. Lynn. An introduction to the analysis and processing of signals. Macmillan, 1973.

2. B.P. Lathi. An introduction to random signals and communication theory. Intertext, London, 1970.

DISCUSSION:

The general directions of the questions fell into three categories: 1) What is the basic question being asked; 2) what are some of the limitations of the method of analysis; and 3) what clinical applications can be expected.

Dr. Pullan noted that the purpose of his research was to determine whether it was possible to quantitate the presence of structure within an image without having to specify its location. Whether this structure or texture can be related to tissue characterization or anatomy was not resolved at this time, primarily because of the limited data base. The results of the research, i.e., the histograms, chi square analysis and auto-correlation data, suggested that structure can be resoved, and it was generally agreed that the method warrants a more comprehensive, statistically adequate study to explore the utility of the method.

The limitations and restrictions of the method relate to the algorithms used. Auto-correlation has an anisotropic condition which makes it dependent upon the direction the data is averaged. This may be advantageous or disadvantageous, depending upon whether one wishes to utilize directional properties. In general, averaging over all directions gives the most consistent data results. Chi square analysis relates to the comparing of differences (means, standard deviations, skewness coefficients) between histograms. It is dependent only upon the assumption that the population of differences are normally distributed. All the methods are very dependent upon machine induced artifacts, such as streaking. They are, however, relatively independent of image noise. Caution is necessitated, however, when comparing similar regions from different images, because the noise on the CT image is positionally dependent, thus requiring that similar anatomical regions be compared. Finally, cross-correlation between normal data and pathologic data gives essentially the same information as comparing the auto-correlation of each data set individually.

The clinical applications to which this work was addressed are threefold: (1) to use the method as a screening procedure to separate normal from abnormal; (2) to use the data in a predictive capacity (i.e., a subtle abnormality is detected on the image which then requires further tests be done so as to elucidate it); and (3) to use the data as a simple descriptor of tissue pathology and possibly tissue type (i.e., a hepatic fibrosis). This preliminary study by Dr. Pullan indicated these are real possibilities which require further investigative efforts.

FULLY-THREE-DIMENSIONAL IMAGE RECONSTRUCTION USING SERIES EXPANSION METHODS

Martin D. Altschuler and Gabor T. Herman
Medical Image Processing Group
Department of Computer Science
State University of New York at Buffalo
Amherst, N.Y. 14226 USA

ABSTRACT

Three-dimensional image reconstruction in tomographic examination is presently achieved by reconstructing a set of sequential transverse sections. Such an approach has a number of adverse consequences, which are absent if truly three-dimensional reconstruction is applied to data derived from a single circumferential scan by a cone-beam x-ray source. We demonstrate that with a careful choice of basis functions, coordinate systems, and matrix indexing, it is possible to achieve truly three-dimensional reconstruction with computer time and memory requirements only slightly in excess of those presently needed to reconstruct and assemble a stack of two-dimensional transverse sections.

INTRODUCTION

Present day CT (computed tomography) scanners reconstruct one transverse section of a subject at a time. If a set of adjacent transverse sections are needed for the examination of a three-dimensional volume of a subject, then present CT scanners must be repeatedly re-positioned.

Section-by-section reconstruction has been successful for a number of reasons. First, the rays used in the reconstruction traverse only the region (section) of interest, so that regions not of interest do not affect the reconstruction. Second, the calculations remain the same for each transverse section because

the geometry of the apparatus is fixed; only the input (x-ray projection) data changes in value from section to section. Third, the methods are simple enough and the numerical data arrays are small enough that the reconstruction can be handled by minicomputers.

Although three-dimensional image reconstruction of an entire volume can be achieved with section-by-section analysis by assembling a stack of transverse sections, the use of this method to examine living organisms and/or dynamic organs is seriously limited by factors of time and axial resolution. The reconstruction of a three-dimensional distribution from a stack of sections may require a considerable number of slices to achieve adequate axial resolution. During the time required to collect data for a number of consecutive sections, dynamic organs such as the lungs change in shape, thereby jeopardizing the registration (axial correlation) of the stack of slices. Synchronization of the x-ray projection data to one particular phase in the cycle of a dynamic organ provides an expedient solution but assumes (not correctly) that a dynamic organ returns to precisely the same state from one cycle to the next. Even apart from the case of dynamic organs, data collection over a few minutes is a considerable time to restrain a patient with the degree of immobility necessary for perfect axial registration of a stack of sections (particularly if the person is young or elderly). Multi-source tomographic devices which obtain several sections simultaneously at different positions along the axis would improve the possibility of successful three-dimensional reconstruction, but (1) are expensive, and (2) require proper mathematical correction for sections lying between those containing the x-ray sources.

In this paper we examine the mathematical problem of reconstructing a three-dimensional distribution using projection data taken with a number of suitably chosen <u>cone-beam</u> x-ray sources located outside the object of interest. Our purpose is to demonstrate that fully-three-dimensional image reconstruction from cone-beam sources can be effected with only a minimal increase in algorithmic complexity beyond what presently exists in computed section-by-section tomographic reconstruction.

Because our algorithms reconstruct volume distributions from cone-beam projections, data pertinent to many transverse sections of a subject can be collected in a single scan (that is, in a few

seconds or less) thereby minimizing the time a patient need be restrained from motion. The methods outlined here are directly applicable to the Dynamic Spatial Reconstructor (DSR), now being built at the Mayo Clinic[1], which can collect in 10 milliseconds all the data necessary for three-dimensional reconstruction.

In cone-beam reconstruction there is considerable freedom in choosing the geometrical positions of the x-ray sources relative to the object. One can for example place the x-ray sources (1) at the vertices of a polyhedron enclosing the object, (2) at equal azimuths on a circle encompassing the object, (3) at equal azimuths around two or more congruent, coaxial, non-coplanar circles encompassing and sandwiching the region of interest, (4) at various distances along selected lines diverging from the object, (5) at random positions on a sphere enclosing the object, etc. The first three geometries have a high degree of symmetry. In general, symmetry in the positioning of the cone-beam sources permits decomposition of the mathematically complex three-dimensional reconstruction problem into a number of independent, mathematically simpler problems.

In this paper we examine in detail the special case of cone-beam sources positioned at equal intervals of azimuth around a circle whose plane is perpendicular to the chosen rotation axis and whose radius is large enough to encompass the object (that is, geometry #2 above). This geometry permits a particularly efficient algorithm for three-dimensional image reconstruction.

If in each projection the object of interest is entirely contained within the cone beam and if no objects not of interest contribute to the projection data, then necessary and sufficient information for three-dimensional image reconstruction is potentially available. For many practical applications, the task is to reconstruct a limited non-isolated region (of a thorax say) which is connected above and below to other parts of the body. In this case, there exist some diverging rays which in addition to passing through the object of interest also traverse regions not of interest. If these divergent rays are included in the reconstruction analysis, they introduce some intrinsic noise to the cone-beam reconstruction problem; if these divergent rays are omitted from the reconstruction analysis, then some essential information concerning

the region of interest is lost. This difficulty in using divergent cone-beam x-rays for three-dimensional reconstruction of non-isolated objects can be greatly mitigated by using geometry #3 above, that is, putting the cone-beam sources on at least two congruent, coaxial, non-coplanar circles and sandwiching the region of interest between them. Three-dimensional image reconstruction with geometry #3 will be the subject of another paper. It is slightly more complicated than the problem discussed here of placing cone-beam sources on only a single circle (geometry #2).

In the approach of this paper, the geometry for the relative positions of the x-ray sources and the object is first translated into matrices; the inverse matrices are then determined and stored in a computer. All subsequent three-dimensional image reconstructions (for that chosen geometry) are achieved by multiplying the inverse matrices with the data vectors. Alternatively, the matrix equations may be solved using iterative methods[2].

In reconstructing three-dimensional volumes M elements on a side, one might expect matrices of order $M^3 \times M^3$. Matrices of such magnitude are clearly forbidding if M exceeds even 30. However, with equal azimuthal spacing of the x-ray sources and with the unknown variables indexed in a particular order, we can reduce the large geometry matrix to a relatively simple form. Because the work required to invert the geometry matrix can be so tremendously reduced by exploiting symmetry properties and careful indexing schemes, we believe fully-three-dimensional image reconstruction can be commercially feasible. Only slightly more computing time should be required for fully-three-dimensional reconstruction than is needed presently to analyze a three-dimensional volume by assembling a stack of transverse sections.

THE MATHEMATICAL PROBLEM

The problem of fully-three-dimensional image reconstruction can be described as follows. We are given (1) an unknown density distribution (actually an unknown distribution of the linear attenuation coefficient) within a finite three-dimensional volume

and (2) cone-beam x-ray sources at positions outside the volume of interest. The cone-beam sources provide two-dimensional projections of the unknown three-dimensional density distribution. The problem is to estimate the unknown distribution of matter (that is, the distribution of the linear attenuation coefficient) from all the available projection data.

Mathematically, the problem can be idealized as follows. Let f be a function of three variables whose value is known to be zero outside a compact set S. Assume that we know the integrals of f along a number of lines. Our problem is to estimate f from these line integrals.

In practice, one can only estimate the line integrals from the x-ray measurements. How such estimations are done is discussed in many places[3]. Our method of solution partially compensates for the errors which arise in estimating the line integrals from the measurements.

Our basic strategy is as follows. We express the unknown density distribution as a series expansion in terms of suitable basis functions and apply a least squares criterion to set up a matrix equation for the unknown coefficients. If we have carefully chosen the geometrical arrangement of the cone-beam sources and the set of basis functions, then the large least-squares matrix can be reduced by suitable indexing to a set of matrices each of which is small enough and simple enough to be inverted by present machines. Once the smaller matrices are correctly inverted and the inverse matrices stored, subsequent fully-three-dimensional image reconstruction reduces to the problem of multiplying the smaller inverse matrices with different data vectors.

THE SERIES EXPANSION APPROACH

We assume that any three-dimensional distribution in a finite volume (that is, with finite support) can be sufficiently approximated by a linear combination of some pre-determined basis functions. The unknown coefficients of this linear combination are

estimated from the equations which result when the projections of the unknown distribution are expressed as linear combinations of the projections of the basis functions. The details of this approach appear elsewhere[2]. Here we give only a summary.

Let the set of basis functions be $\{f_1(\vec{r}),\ldots,f_I(\vec{r})\}$. In the series expansion approach, the problem of determining the unknown three-dimensional function $f(\vec{r})$ is replaced by the problem of determining a finite number of coefficients $c_1, \ldots, c_I$ such that $\hat{f} = \sum_{i=1}^{I} c_i f_i$ is an "acceptable" approximation to f. We shall define "acceptability" in the next section by requiring that f is as consistent with the data as possible in the least squares sense.

We assume that we made measurements to obtain the value of the line integral of the unknown function f along N rays. For $1 \leq n \leq N$, let $g(n)$ denote (our estimate, based on measurement, of) the line integral of f along the nth ray, and, for $1 \leq i \leq I$, let $g_i(n)$ denote the line integral of f_i along the nth ray. Then,

$$f \approx \sum_{i=1}^{I} c_i f_i \tag{1}$$

implies[2] that for all n

$$g(n) \approx \sum_{i=1}^{I} c_i g_i(n). \tag{2}$$

Since the values of $g(n)$ are measured and the values of $g_i(n)$ can be calculated, the equations (2) can be used to estimate the values of the c_i, which then can be substituted into equation (1) to estimate f.

Usually there will not exist any set $c_1, \ldots, c_I$ such that the two sides of equations (2) match exactly for all n. On the other hand, sometimes there is more than one set. There are many different criteria which lead to a unique set $c_1, \ldots, c_I$ (see ref. 2); in this paper we describe the least-squares criterion.

We minimize

$$D = \sum_{n=1}^{N} [g(n) - \sum_{i=1}^{I} c_i g_i(n)]^2.$$

The I unknown coefficients $\{c_i\}$ which minimize D satisfy the set of I equations

$$\sum_{n=1}^{N} g(n) g_j(n) = \sum_{n=1}^{N} g_j(n) \sum_{i=1}^{I} c_i g_i(n), \qquad 1 \leq j \leq I.$$

The I linear equations in I unknowns can be rewritten in matrix form. Defining the matrices G, A, and C in component form by

$$G_j = \sum_{n=1}^{N} g(n) g_j(n),$$

$$A_{ij} = A_{ji} = \sum_{n=1}^{N} g_i(n) g_j(n),$$

$$C_i = c_i,$$

we can write the I linear equations as

$$G = A\ C.$$

If A is non-singular, the C array is given for the least-squares solution by

$$C = A^{-1}\ G.$$

Once we know C, we can construct the approximation of f.

We note in passing that calculation of the components of both G and A requires evaluation of $g_i(n)$ for $1 \leq i \leq I$ and $1 \leq n \leq N$. Since each $g_i(n)$ is the integral of a basis function along a line, and I and N are typically large (for example, of the order 10^4 to 10^7), we need to perform a very large number of line integrals. However, by carefully choosing a reference frame which locates points relative to the rays emanating from a given source position and then using the necessary orthogonal transformations to get to and from the absolute reference frame, we can perform these integrations with surprising ease. In addition, (as we shall show in the penultimate section), for a carefully chosen

geometry, most of these integrals can be expressed in terms of the others, and need not be evaluated by explicit integration.

Another comment is that when we talk about the inversion of matrices and storing their inverses, we mean this in a most general sense. In practice, it is likely to be worthwhile to use triangular decomposition of the matrix A rather than explicit inversion. It will be clear throughout our work that the savings which our methods provide for a particular inversion method carry over to any other method for solving $G = A\ C$.

In the remainder of this paper we shall assume that the basis functions and the rays are chosen in such a way that the matrix A defined above is invertible. The inverse need only be found once (for any particular choice of the basis functions and rays); to obtain C for a given set of measurements we need only multiply the "data" vector G by A^{-1}. Hence, the user of this method incurs a one time cost of inverting the matrix A and a repeated cost of multiplying with A^{-1}.

A serious difficulty arises if the matrix A is large. This is because the inversion (say by Gaussian elimination) of an $I \times I$ matrix requires of order $I^3/3$ computer multiplications, and multiplying a vector by an $I \times I$ matrix requires I^2 computer multiplications. In our problem the size of the matrix A is $I \times I$, where I is the number of basis functions. When the reconstructed images, which are always linear combinations of basis functions, are to be used for detailed diagnostic examination, I usually must be large, typically 10^4 to 10^7. Obviously, the one time cost of $I^3/3$ multiplications would put an impossible burden on any present generation computer. The major thrust of our paper, therefore, is to discuss how the basis functions and the rays can be chosen to produce a matrix A of such structure that the computational problems just mentioned can be avoided.

Suppose that there exists an integer L such that the matrix A can be decomposed into submatrices as follows

$$A \doteq \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,L} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,L} \\ \vdots & \vdots & & \vdots \\ B_{L,1} & B_{L,2} & \cdots & B_{L,L} \end{bmatrix}$$

where $B_{i,j}$ is a matrix for $1 \leq i \leq L$ and $1 \leq j \leq L$. Suppose further that there exists an e such that $B_{i,j}$ is the null matrix if $|i - j| > e$. In such a case, A is called a <u>block polydiagonal matrix</u>. Provided e is much smaller than L, the inversion of such matrices (as well as the multiplication of a vector by the inverse) is much less expensive on a computer than the estimates based on the size of A would indicate[4].

For example, if $e = 0$ (that is, A is a block-diagonal matrix) then

$$A^{-1} = \begin{bmatrix} B_{1,1}^{-1} & 0 & \cdots & 0 \\ 0 & B_{2,2}^{-1} & \cdots & 0 \\ & & \ddots & \vdots \\ & & & 0 \\ 0 & 0 & \cdots & B_{L,L}^{-1} \end{bmatrix}$$

Assuming that each of the matrices $B_{i,i}$ are of the same size, the inversion of A requires $L(1/3)(I/L)^3 = (1/3)I^3/L^2$ computer multiplications, and multiplication of a vector by the inverse of A requires $L(I/L)^2 = I^2/L$ computer multiplications. Since in our applications we expect L to lie between 30 and 100, the inversion of such a block diagonal matrix is 10^3 to 10^4 times faster, and multiplication by such a matrix is 30 to 100 times faster, than the corresponding operations for full matrices of the same dimension.

In what follows we shall try to produce matrices A which are block polydiagonal with L as large as possible (greater decomposability) and e as small as possible (more nearly block diagonal).

As a conceptually, but not computationally, simple special case we first consider basis functions f_i corresponding to voxels (volume elements). These are defined by

$$f_i(x, y, z) = X_k(x)\ Y_\ell(y)\ Z_m(z)$$

where k, ℓ, m, are integers in the domain [1, 2M-1], and i is the index $i = k + [(\ell-1) + (m-1)(2M-1)](2M-1)$. The functions X_k, Y_ℓ, and Z_m are defined by

$$X_k(x) = \begin{cases} 1, & \text{if } |x - M + k| \leq \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

$$Y_\ell(y) = \begin{cases} 1, & \text{if } |y - M + \ell| \leq \frac{1}{2}, \\ 0, & \text{otherwise} \end{cases}$$

$$Z_m(z) = \begin{cases} 1, & \text{if } |z - M + m| \leq \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus the value of each basis function is unity within some unit cube (centered at a point with integer coordinates) and 0 outside. Combined together, these cubes cover a volume which is a cube whose sides are of length 2M. The choice for the index i arranges the basis functions into a linear array of size $I = (2M-1)^3$ in the order of "slices" (m), "rows" (ℓ), and "columns" (k), with slices changing slowest and columns the fastest.

Using the notation of the last two sections we may define, for $1 \leq n \leq N$,

$$U_{k,\ell,m}(n) = g_i(n),$$

where i is determined from k, ℓ, and m by the formula given above. Then,

$$G_i = \sum_{n=1}^{N} g(n)\ U_{k,\ell,m}(n)$$

and, provided $j = k' + [(\ell'-1) + (m'-1)(2M-1)](2M-1)$,

$$A_{ij} = A_{ji} = \sum_{n=1}^{N} U_{k,\ell,m}(n)\ U_{k',\ell',m'}(n).$$

The size of the matrix A is $(2M-1)^3 \times (2M-1)^3$. Since in typical tomographic applications the size of M may well exceed 100, the least squares analysis of the last section requires inverting a matrix whose size may be $10^7 \times 10^7$, or even greater. Since straightforward matrix inversion techniques are not feasible for matrices of such size, the practicality of this approach to fully-three-dimensional image reconstruction depends on one being able to reduce the magnitude of the calculations involved in inverting the matrix A.

Consider, first of all, a pair of voxels i and j (as defined above). If there is not any ray n which traverses both voxels i and j, then at least one of $g_i(n)$ and $g_j(n)$ is zero and the matrix element $A_{i,j}$ is zero.

Consider now the special case in which for every ray there is a slice such that the ray lies entirely within the slice; different rays may lie in different slices. (This may occur, for example, if the source positions are in the z=0 plane far from the object to be reconstructed.) In this case, given any pair of voxels i and j such that voxel i and voxel j are in different slices (i.e., have values m_i and m_j of the index m with $m_i \neq m_j$), there is no ray traversing both of the voxels and hence $A_{i,j} = 0$. The indexing function defined above now ensures that the matrix A is block diagonal with adjacent blocks corresponding to adjacent slices and with the number of blocks L equal to the number of slices 2M-1. Thus the three-dimensional reconstruction in this case is equivalent to a slice-by-slice reconstruction.

More generally, if all the source positions are in the z=0 plane, the matrix A become block polydiagonal, with e equal to the maximum number of slices which is traversed by any single ray. (This happens because our indexing function changes slowest with the slices.) If the source is kept at a moderate distance from the object, e will be small (typically less than 10) and inversion of the matrix A is made feasible by techniques developed for polydiagonal matrices[4].

However, one can do much better than this by choosing the basis functions and source positions more carefully.

ANALYSIS IN TERMS OF CYLINDRICAL COORDINATES

In this section we use cylindrical coordinates: r(radius), t(azimuth), and z(height). We assume that all the source positions lie on a circle in the plane z=0 placed at equal azimuthal intervals. The azimuths of the source positions are $2p\pi/P$ for $0 \leq p < P$. Furthermore, we shall assume that the set of rays along which the integral of the unknown function f is measured is invariant under rotation by the angle $2\pi/P$ around the axis $r = 0$. This is a reasonable assumption; it is satisfied by any apparatus where a single source and a set of detectors form a rigid object which rotates around the $r = 0$ axis in equal angular increments. We shall also assume that for each of the P source positions there are Q rays along which the integrals of f are measured. Thus the total number of rays is $N = P\,Q$.

Let d_r and d_z be positive real numbers (we may think of them as the desired radial and height resolutions). Let K, L, and M be positive integers. Define the basis functions by

$$f_i = R_k(r)\ T_\ell(t)\ Z_m(z),$$

where k, ℓ, m are integers in domains $[1,K]$, $[0,2L]$, and $[1,2M-1]$ respectively and i is the index

$$i = k + [(m-1) + \ell(2M-1)]K.$$

The functions R_k, T_ℓ, Z_m are defined by

$$R_k(r) = \begin{cases} 1, & \text{if } |r - r_k| \leq d_r/2, \\ 0, & \text{otherwise,} \end{cases}$$

$$T_\ell(t) = \begin{cases} \cos\left(\frac{1}{2}\ell t\right), & \text{if } \ell \text{ is even,} \\ \sin\left(\frac{1}{2}(\ell+1)t\right), & \text{if } \ell \text{ is odd,} \end{cases}$$

$$Z_m(z) = \begin{cases} 1, & \text{if } |z-z_m| \leq d_z/2, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$r_k = \left(k - \frac{1}{2}\right) d_r,$$

and

$$z_m = (M - m)\ d_z.$$

Thus, the number of basis functions we use is

$$I = K(2L + 1)(2M - 1).$$

Note that the choice for the index i arranges the basis functions in a linear array in the order of "Fourier components" (ℓ), "slices" (m), and "cylindrical shells" (k), with Fourier components changing most slowly and cylindrical shells changing most rapidly.

We now show that for these basis functions and for this indexing scheme, the structure of the matrix A is particularly simple.

We define, for $1 \leq k \leq K$, $0 \leq \ell \leq 2L$, $1 \leq m \leq 2M-1$, $1 \leq p \leq P$, and $1 \leq q \leq Q$,

$$U_{k,\ell,m}(p,q) = g_i(n),$$

where i is as defined above and

$$n = q + (p-1)Q.$$

Then, provided

$$j = k' + [(m'-1) + \ell'(2M-1)]K,$$

we have matrix elements

$$A_{i,j} = A_{j,i} = \sum_{q=1}^{Q} \sum_{p=1}^{P} U_{k,\ell,m}(p,q)\ U_{k',\ell',m'}(p,q)$$

First we show that provided $P > 2L$ we have $A_{i,j} = A_{j,i} = 0$, unless either $\ell = \ell'$ or $|\ell'-\ell| = 1$ with the larger of ℓ,ℓ' being even. Thus A is a block polydiagonal matrix in which the first of the L + 1 diagonal blocks is of size K(2M-1) and the rest are of size 2K(2M-1). The proof makes use of two lemmas (the proofs of which are omitted). For $1 \leq p \leq P$, we use t_p to abbreviate $2\pi(p-1)/P$.

<u>Lemma 1</u>. For any k, ℓ, m, p, and q within the specified ranges, if ℓ is even, then

$$U_{k,\ell,m}(p,q) = \cos(\tfrac{1}{2}\ell\ t_p)U_{k,\ell,m}(1,q) - \sin(\tfrac{1}{2}\ell\ t_p)U_{k,\ell-1,m}(1,q)$$

(where the second term is zero if $\ell = 0$), and if ℓ is odd, then

$$U_{k,\ell,m}(p,q) = \sin(\tfrac{1}{2}(\ell+1)t_p)\ U_{k,\ell,m}(1,q) + \cos(\tfrac{1}{2}(\ell+1)t_p)\ U_{k,\ell+1,m}(1,q).$$

<u>Lemma 2</u>. If u and v are non-negative integers such that $0 < u+v < P$, then

$$\sum_{p=1}^{P} \cos u\, t_p \sin v\, t_p = 0,$$

and, provided also that $u \neq v$,

$$\sum_{p=1}^{P} \cos u\, t_p \cos v\, t_p = 0,$$

$$\sum_{p=1}^{P} \sin u\, t_p \sin v\, t_p = 0.$$

From these two lemmas our claim easily follows. For example, if $\ell = 4$ and $\ell' = 9$, then

$$\begin{aligned}
A_{ij} &= \sum_{q=1}^{Q} \sum_{p=1}^{P} U_{k,4,m}(p,q)\ U_{k',9,m'}(p,q) \\
&= \sum_{q=1}^{Q} U_{k,4,m}(1,q)\ U_{k',9,m'}(1,q) \sum_{p=1}^{P} \cos 2t_p \sin 5t_p \\
&\quad + \sum_{q=1}^{Q} U_{k,4,m}(1,q)\ U_{k',10,m'}(1,q) \sum_{p=1}^{P} \cos 2t_p \cos 5t_p \\
&\quad - \sum_{q=1}^{Q} U_{k,3,m}(1,q)\ U_{k',9,m'}(1,q) \sum_{p=1}^{P} \sin 2t_p \sin 5t_p \\
&\quad - \sum_{q=1}^{Q} U_{k,3,m}(1,q)\ U_{k',10,m'}(1,q) \sum_{p=1}^{P} \sin 2t_p \cos 5t_p \\
&= 0.
\end{aligned}$$

We have now reduced the matrix A to block-diagonal form with L + 1 blocks. Thus the inversion of A can be carried out block by block, with considerable savings in computer time and memory requirements.

Further simplification can be achieved in handling the individual blocks on the diagonal of the matrix A. Within any of these blocks, the indexing scheme we have described has the values of m (slices) changing slower than the values of k (cylindrical shells). Following the arguments presented in the previous section, we find that each of the L + 1 blocks on the diagonal of A is itself a block polydiagonal matrix, with usually a small value of e. Thus the inversion of the blocks of the matrix A is a considerably simpler matter than their sizes would indicate.

Note that Lemma 1 also gives an indication of the validity of our earlier claim, that, in practice, the number of explicit integrations required to calculate the $g_i(n) = U_{k,\ell,m}(p,q)$ is much fewer than N I.

CONCLUSION

We have demonstrated that the amount of calculation needed for fully-three-dimensional image reconstruction can be reduced to manageable size when data from cone-beam x-ray sources equally spaced in azimuth are used together with a carefully designed indexing scheme for particular basis functions in cylindrical coordinates.

ACKNOWLEDGEMENTS

The authors are grateful to members of the Medical Image Processing Group at the State University of New York at Buffalo, especially to Drs. Arnold Lent and Donald Solmon, for many helpful discussions. Research for this work is supported by NIH grants HL 18968, 4664, and GM 381, NCI contract CB 53860 and NSF grant MCS75-22347.

REFERENCES

1. E. L. Ritman, R. A. Robb, S. A. Johnson, P. A. Chevalier, B. K. Gilbert, J. F. Greenleaf, R. E. Sturm, and E. H. Wood, "Quantitative Transaxial Imaging of the Anatomic Structure and Function of the Heart, Lungs, and Circulation", Mayo Clinic Proc. (in press).

2. G. T. Herman, and A. Lent, "Iterative Reconstruction Algorithms", Comput. Biol. Med. 6, 273 (1976).

3. G. T. Herman, Correction for Beam Hardening in Computed Tomography, State University of New York at Buffalo, Department of Computer Science Technical Report Number 129 (June 1977).

4. E. Isaacson and H. B. Keller, Analysis of Numerical Methods, John Wiley & Sons, Inc., New York, N.Y., 1966.

DISCUSSION:

Pullan asked Altschuler if he could estimate the time required to obtain a three-dimensional reconstruction if one processes cone-beam data from,say,a 160 x 160 detector array. Altschuler noted that the time required for processing cone-beam data into a three-dimensional reconstruction will probably be slightly in excess of what is now required to process a series of transverse sections into a three-dimensional reconstruction. The patient confinement time needed to collect projection data for three-dimensional reconstruction will, of course, be greatly reduced with cone-beam apparatus.

Another participant asked Altschuler if he would discuss the problems that scatter would introduce in terms of the coupling terms between the diagonal matrix blocks.

Altschuler stated that he would not like to talk about the effects of scatter prior to the computer implementation of the algorithm. He added that when the implementation of the algorithm is completed, they plan to evaluate reconstructions obtained from both real noisy data and artificial noise-free data.

Oppenheim asked if the reason he wanted two source rings was to obtain a complete 360°. Altschuler said no and added that the primary reason for putting sources on two congruent coaxial non-coplanar circles with the object of interest sandwiched between is to avoid contamination of the data. Data contamination occurs when a ray passes through matter outside the region of interest, thus introducing an unknown error into the projection measurement. One would like to choose the source-object geometry so that data contamination is completely avoided.

Oppenheim asked if it were correct that one can't take all possible rays from each of the rings. Altschuler noted that that was correct and added that if one had two source circles, one would limit the domain of angles that would be used. However, one would then be able to reconstruct the volume of interest, or at least most of it, without fear of data contamination.

Oppenheim asked if he were limited to only one source ring, what could one do. Altschuler stated that one would have to accept a certain amount of data contamination. How bad it would be depends on many factors. Clearly, if one places the region of interest within (or as close as possible to) the plane of the source circle or makes the radius of the source circle very large relative to the object, the contamination problem should not be serious. If the object to be reconstructed is sufficiently small that it lies completely within the cone beam, then there is no contamination at all. For thorax reconstruction, however, those rays making a large angle to the plane of the source circle (that is, those rays which intersect many transverse sections) might be seriously contaminated.

Another participant asked if one had equally spaced detectors about a circle, would there be problems with the divergence of the matrices?

Altschuler noted that this is a least-squares matrix and added that, with enough data values relative to the number of unknowns, it should be non-singular. In fact, with equal azimuthal spacing the matrix should be well-conditioned for the situation he described.

Expanding upon the previous question, another participant asked if the non-singular nature of the matrix depended upon the particular choice of spacing the detectors.

Altschuler noted that with a poor choice of source positions, one can make the matrix very ill-conditioned; and with insufficient data relative to the number and types of unknowns, the matrix can be made singular. But, in general, one can place the detectors anywhere in three-dimensional space. The price one pays for that freedom is the loss of symmetry. The less symmetry one has, the more difficult it is to decompose the enormous least-squares matrix into diagonal blocks (sub-matrices). One would like to design the geometry and the algorithm so that matrix inversion is possible. He added that if one has a big enough computer and enough time, one can invert bigger matrices and not worry as much about what geometry is used.

Pizer noted that Altschuler mentioned a number of times that he planned to invert the matrices. Pizer added that the standard technique for solving equations for his situation is a triangular factorization approach rather than an inversion approach. Pizer asked if there was any reason why that technique was not applicable here.

Altschuler responded by saying he has just begun the implementation stage of the problem and added that they are now searching for ways of treating block polydiagonal matrices and sparse matrices in general. He stated that they hoped to test a number of subroutines in the software.

Pullan noted that they have thought of alternative ways of minimizing the reconstruction problem. In the full three-dimensional problem one has to reduce the problem as was stated. When one looks at a CT scan one finds it is characterized by areas of high uniformity which are unordered, such as liver, spleen, or kidney. Pullan then asked Altschuler if he had ever considered just reconstructing the edges between these regions, noting that he would

then reconstruct on the basis of parallax between two views taken very close to each other in angle. Any shifts in position of an edge should provide data to allow one to reconstruct surfaces.

Altschuler stated that he had and noted that if one has an edge or a surface one can use parallax. But in the case of walking in a "fog" as one would be doing in three-dimensional tomography, it is very difficult to see parallax.

Pullan responded by stating that we're not looking into a fog but are looking into a highly structured object. He further questioned if he had thought of using the known features of those structures to help do the reconstruction.

Altschuler responded by saying that if one can locate the surfaces a priori it might work. The edges between bone and tissue might be readily obtainable, but prior to reconstruction he didn't know how well one could delineate edges between organs that differ by only a few percent in density.

Coulam suggested that part of the solution to this problem might be in the ultimate display of the three-dimensional data that one chooses and noted that Herman was going to talk about that sort of thing later on.

Altschuler noted that in Dr. Herman's talk one certainly sees boundary detection and ways of pulling out the heart in three-dimensions.

Ashare asked Altschuler if he wanted to explain how one detects edges when doing reconstruction. Altschular stated that he could not detect boundaries prior to reconstruction.

Pullan stated that the edges exist in projections if they exist in three dimensions. Pullan added that it was an approach to reconstruction which he had never seen implemented; and as the speaker obviously thought a lot about ways of contracting the problem, he was asking how he felt about using information one already has about the object in reconstruction. It's the edges that carry the information.

Ashare responded to Pullan by saying that it was rather obvious that he was doing reconstruction by slices rather than by cone beams. In a slice one definitely has the information from an entire slice that can be applied to the next slice. But that's not what Altschuler is suggesting. He is suggesting a cone beam obtaining a three-dimensional picture all at once. Ashare stated that it was difficult for him to comprehend how one would look at edges before one knows that they are there.

Altschuler agreed and added that the way he has seen most things implemented is that first one does the reconstruction, then looks for the edges. What Dr. Pullan is suggesting is that one uses some a priori information directly in the reconstruction. Maybe these comments would be more appropriate after Dr. Herman's talk which concerns boundary detection.

COMPARISON OF RADIOGRAPHY AND COMPUTED TOMOGRAPHY

D. J. Goodenough, K. E. Weaver and D. O. Davis
George Washington University Medical Center
Washington, D.C.

ABSTRACT

The overall assessment of diagnostic image quality reflects a complicated interrelationship of many physical and psychophysical parameters. An examination is made of the relative image parameters of screen/film radiography and computed tomography, concentrating primarily on large area contrast, noise, and signal-to-noise parameters. A relatively simple detection model using known physical image parameters of radiography and computed tomography may be restricted to signals (tumors) on the order of several millimeters to a few centimeters.

Certain parameters are known classically to affect the human ability to visually perceive the existence or absence of some abnormality or "signal" in a diagnostic image. These parameters include contrast, noise, and sharpness.

The classical relationship between the minimal required contrast, the size of the signal, and the brightness of the image can be seen from Figure 1. At normal viewing luminance levels, even in the optimum case of large signals and nominally noiseless backgrounds, one is limited to perceiving signals of about 1% contrast.[1]

The noise in a diagnostic x-ray image is ultimately limited by the finite number of x-ray photons that may be used in making up the image. This finite number of x ray is related to the sensitivity of the diagnostic system used to obtain the image. The amplitude and appearance (character) of the noise fluctuations can diminish the ability to perceive a radiologic signal.

The sharpness of the boundary of the signal from the background is determined by the intrinsic resolution properties of the diagnostic imaging system, as well as motion unsharpness arising from the inherent movement of the anatomical part of interest and the particular exposure time required to record the image.

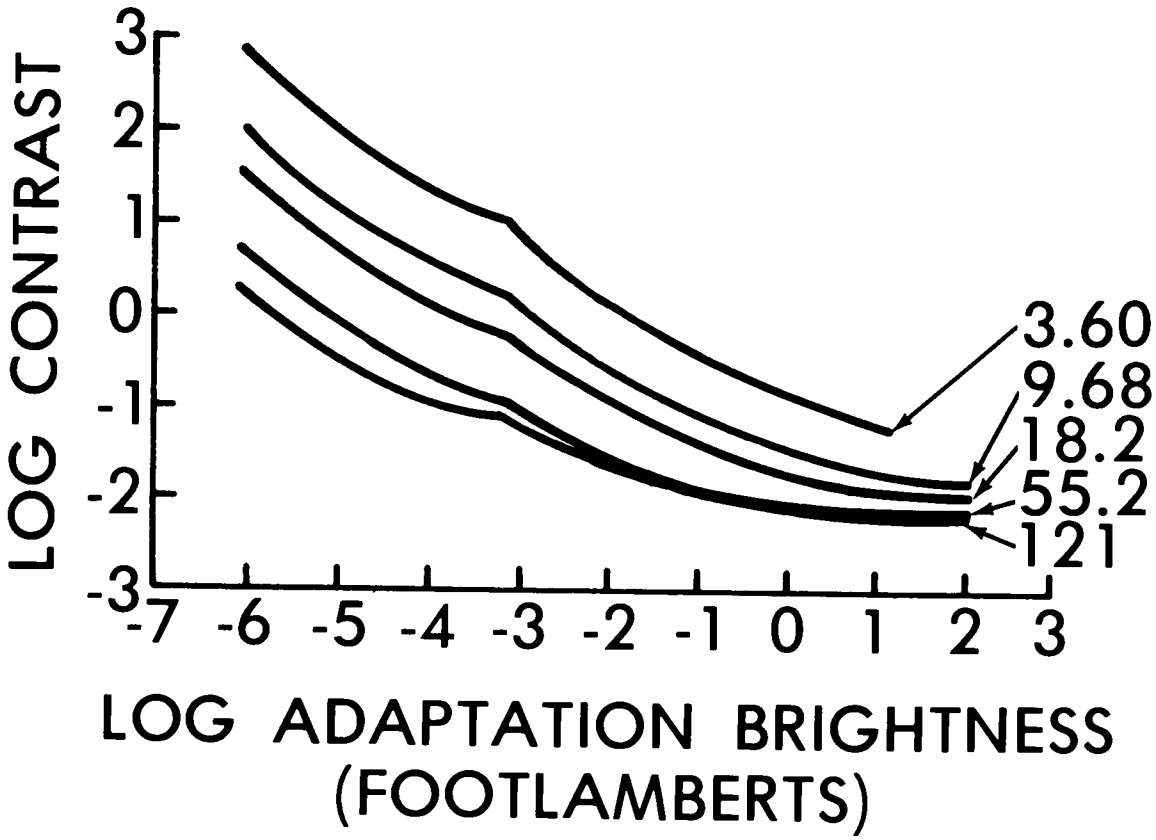


Fig. 1: Relationship between required Contrast, Adaptation brightness, and the size of the signal (plotted as minutes of arc). Adapted from Blackwell.

It should be kept in mind that the overall assessment of diagnostic image quality reflects a complicated interrelationship of many physical and psychophysical parameters, as may be seen from Table I. In addition, the important questions of pattern recognition and structure noise further complicate the evaluation of diagnostic image quality.[2]

TABLE I

Experimental parameters to be controlled and specified in evaluation of human observer detection performance.

(i) Image parameters

- (a) Signal to be detected: size, shape, inherent contrast
- (b) Number of possible signals
- (c) Image system: spatial resolution (MTF), sensitivity, linearity, noise (amplitude and character), speed (dose)
- (d) Non-signal structure: grid lines, anatomy

(ii) Observational parameters

- (a) Display system
  - Brightness scale
  - Gain
  - Offset
  - Any nonlinearity
  - Magnification/minification
- (b) Viewing condition
  - Viewing distance
  - Ambient room brightness
- (c) Detection task
- (d) Number of observations

(iii) Psychological parameters

- (a) *a priori* information given observer
- (b) Feedback given observer (if any)
- (c) Observer experience for other given parameters
  - Clinical versus non-clinical familiarity with signal, display, and especially with type of noise artifacts to be expected.

In this paper, an attempt will be made to examine relative image parameters of screen/film radiography and Computed Tomography. The examination will concentrate primarily on large area contrast, noise, and signal to noise parameters that might be used to describe the limitations on the detection of an isolated tumor-type signal embedded in a homogeneous surround. It is recognized that this model is useful in trying to determine basic physical limitations of the two types of diagnostic imaging systems. (Figure 2)

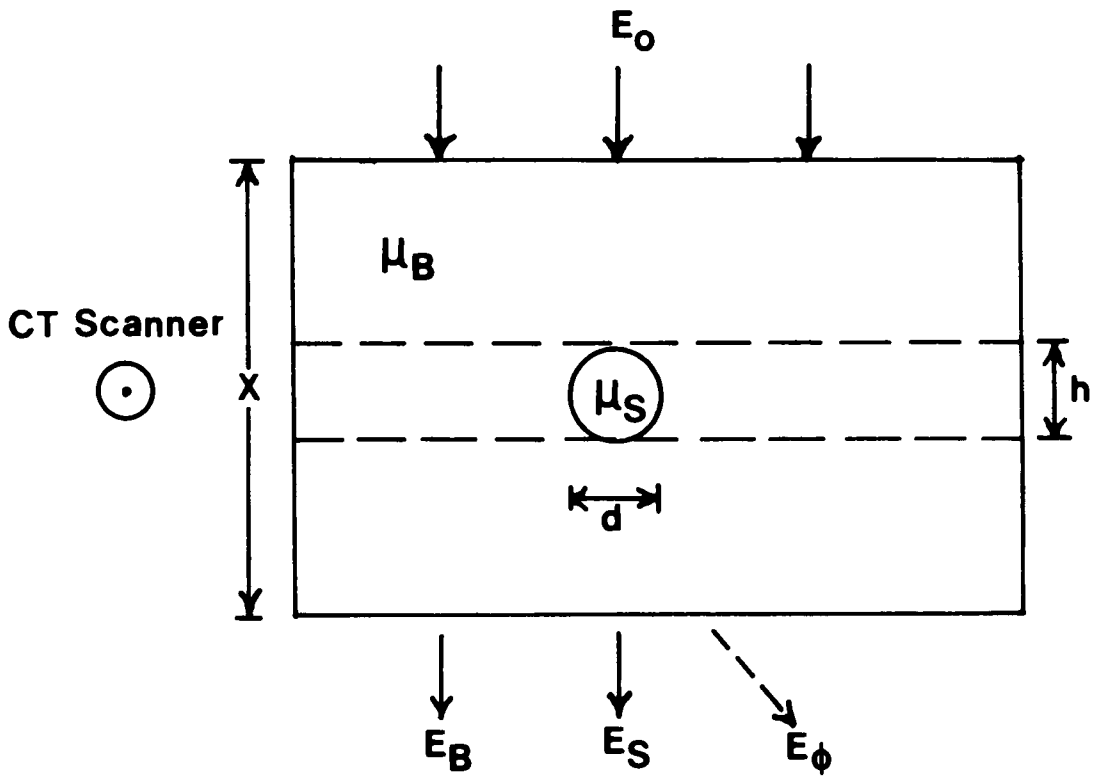


Fig. 2. Schematic illustration of model used to compare radiography and Computerized Tomography. Spherical signal of diameter d and absorption coefficient μ_S imbeded in homogenous media μ_B. E_o is the incident exposure, E_B the exit exposure from the surround, E_S the exit exposure from the surround + signal and E_ϕ the exposure from scattered radiation. The CT slice width is given by h.

CONTRAST IN RADIOGRAPHY AND COMPUTED TOMOGRAPHY

Contrast in Radiography

In considering factors that limit the amount of contrast between a radiographic signal and its surround, the inherent "subject contrast" of the signal should first be examined. Subject contrast is defined as the ratio of the x-ray intensities transmitted by two selected portions of the subject. Subject contrast depends on the physical composition of the subject, the x-ray radiation quality or spectral distribution, and the intensity and distribution of the scattered radiation. "Subject contrast" is modified by the screen/film combination into "radiographic contrast" which is defined as the difference in optical density between two selected portions of a radiograph. The relationship between subject contrast and radiographic contrast is probably best understood by examination of the well-known characteristic curve of a medical x-ray film (The H&D Curve). The characteristic curve of the screen/film depends upon the type of film and the processing conditions. Its shape is usually independent of the quality and distribution of the x-ray radiation reaching the fluorescent screens. An H&D curve is characterized by the relationship between density (D) and logarithm of x-ray exposure (E) reaching the screen/film combination:

$$D = G \log_{10} E + K \qquad \text{(equation 1).}$$

here G is the slope of the approximately straight line portion of the curve between densities of 0.25 to 2.0 above the base plus fog (G is referred to as the average film gradient) and K is a constant.

This equation may be reduced to a convenient working form in that for an average x-ray exposure, $\bar{E}$, it is found that:

$$\Delta D = 0.43G\ \Delta E/\bar{E} \qquad \text{(equation 2)}$$

where small changes in density, ΔD, may be related to small changes in x-ray exposure, ΔE. Moreover, the relationship between density and the transmitted light coming through the radiograph (when placed on the viewbox) is given by:

$$D = -\log_{10} T \qquad \text{(equation 3).}$$

It can be shown (see Appendix 1) that the radiograph contrast, C_R, is approximately equal to

$$C_R = G(\mu_S - \mu_B)d\ \frac{E_B}{E_B + E_\phi} \qquad \text{(equation 4),}$$

where d is the diameter of the signal, E_ϕ is the exposure due to scattered radiation, and E_B is the exposure due to primary radiation. Equation 4 indicates that increased radiographic contrast can be obtained from increasing G, or from increasing difference between the attenuation coefficient of the signal and background, $\mu_S - \mu_B$ (i.e. subject contrast), or by increasing the diameter d of the signal, or by reducing scattered radiation.

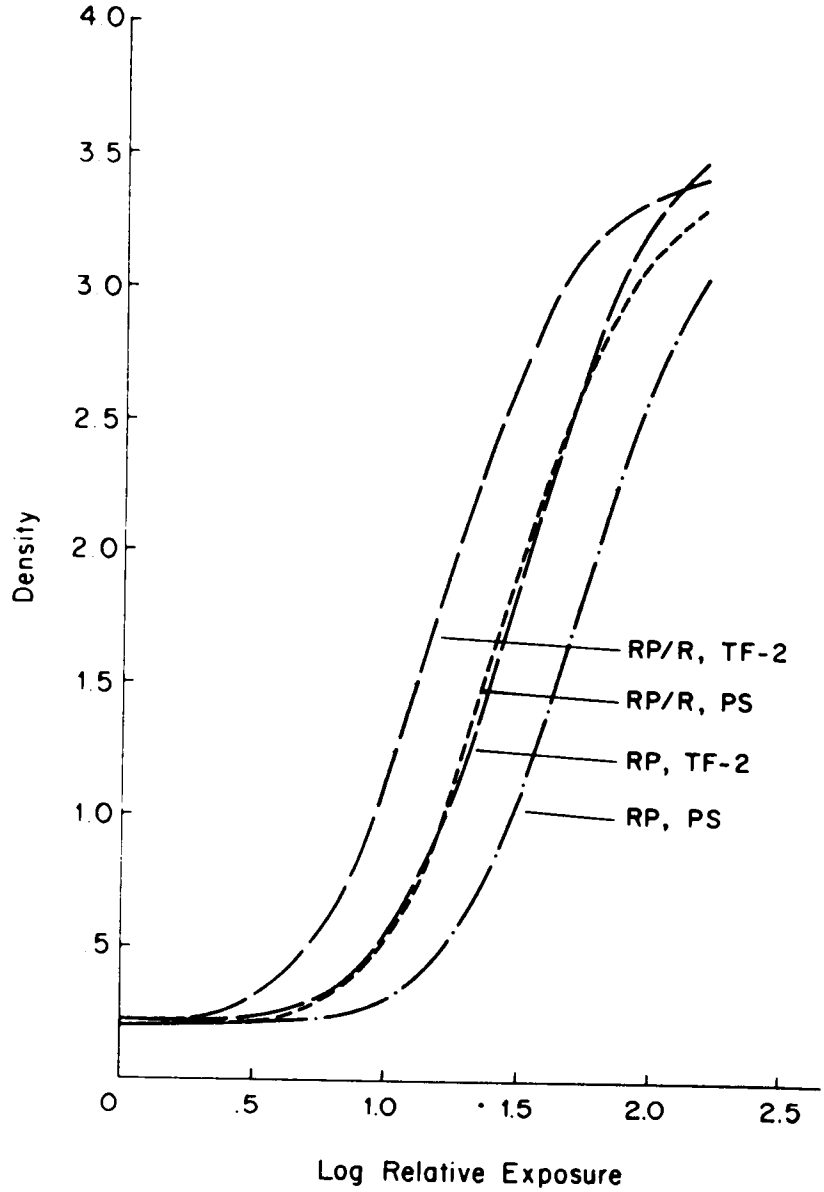
The typical H&D curves shown in Figure 3 are characterized by three major parameters: the density level of the base plus fog that exists in the absence of added x-ray exposure; the average gradient G; and the shoulder region (that is the region in which the film essentially no longer can respond to changes to increased relative x-ray exposure). One should keep in mind the contrast latitude trade-off found in radiographic systems. The higher the slope (G) of an H&D curve, the greater the radiographic contrast as can be seen from equation 2. On the other hand, the latitude, that is the range of exposures (and thus densities) which may be adequately imaged on a single film, is reduced as the slope is increased because of the finite exposure range which may be recorded between the toe region and the shoulder region.

Attenuation coefficients of potential radiologic signals (μ_S) as a function of energy are given in Figure 4. If it is assumed that the background coefficient is a tissue or water-like substance, then, $\mu_S - \mu_B$, (subject contrast), usually decreases as a function of increased energy. As an approximate rule of thumb, the increase in subject contrast can be assumed to follow a power function relationship of kVp^{-n} (where n is approximately 2 in the diagnostic energy range, as can be seen from Figure 4).[*3] The exact dependence of contrast on kVp settings is of course complicated by the fact that the diagnostic x-ray beam is of a polychromatic nature, as shown in Figure 5. In addition, subject contrast is degraded by scattered radiation, which is also a function of energy. Scattered radiation basically tends to irradiate the whole area of the film more or less uniformly, thus producing a general fogging of the film. The net effect of scatter is a reduction in contrast.

The relative abundance of scattered radiation compared to the primary radiation reaching the screen/film increases as the volume or area of the patient irradiated increases and as the kilovoltage (or photon energy) is increased.

Fig. 3.

Characteristic (H&D) curves of four different film-screen combinations.

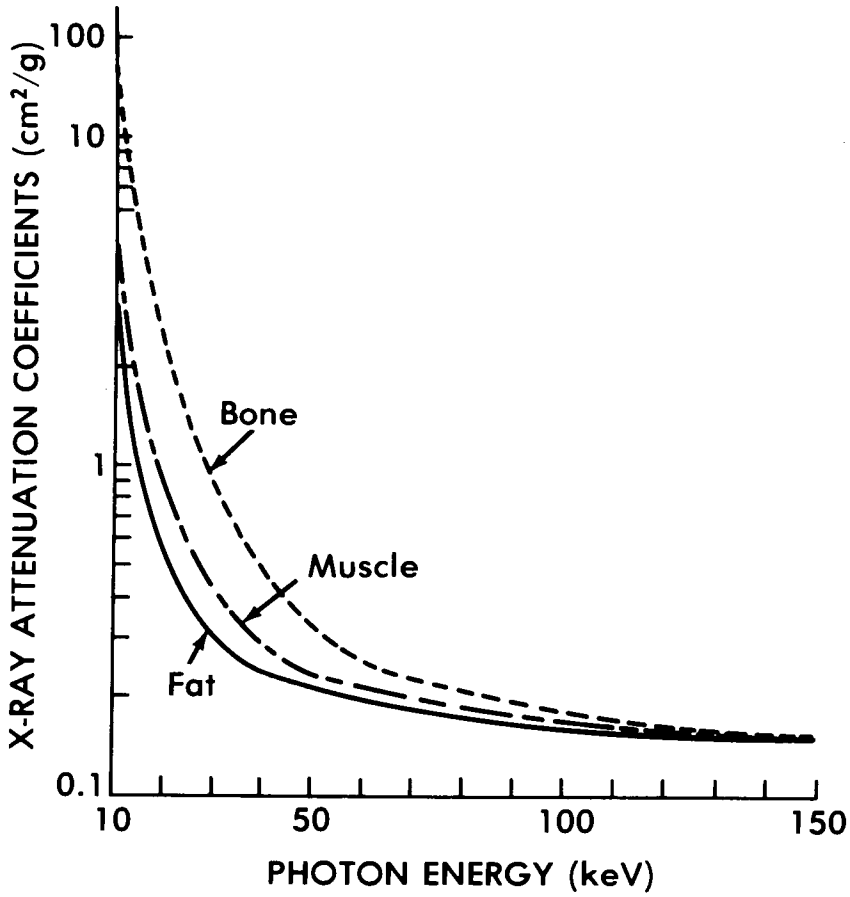


The contribution from scattered radiation in radiography can approach as much as 90% in the case of the abdomen, 50% for the chest, and as much as 25% for extremities such as the knee.[4] The effects of scatter radiation on image quality is shown in the skull radiograph of Figure 6a, where no attempt has been made to remove scattered radiation.

An x-ray grid is the most commonly used device for reducing scattered radiation. An example of the importance of scatter removal (using an x-ray grid) is shown in Figure 6b. The amount of scattered radiation reaching the film can also be reduced by using collimation to limit the useful beam to the area of radiographic interest. An example of collimation is shown in Figure 7, where the radiograph of the skull has been confined to a small area around the sella turcica. Comparison of this radiograph with that in Figure 6a clearly shows the marked increase in radiographic contrast due to scatter reduction.

Fig. 4.

X-ray attenuation coefficients as a function of energy.



Another means of reducing radiographic scatter is by using a lower kVp beam because in the diagnostic energy range, the ratio of scattered to nonscattered radiation usually increases with increasing photon energy; however, a low kVp technique is not always practical due to the thickness of the body parts and the higher patient exposure resulting at lower kilovoltages. Use of an air gap is another means of reducing scattered radiation. In addition, recent progress has been reported in minimizing scatter by using a moving slit assembly to obtain narrow beam geometry in radiographic applications.* [5]

Contrast in Computed Tomography

In the case of homogeneously composed volume elements, the CT output has been reported to be related linearly to the "effective" linear attenuation coefficient (μ), characterizing the volume element under investigation. In fact, it has been reported that the CT number, §, is given by:

$$\S = K_1 \frac{(\mu-\mu_w)}{\mu_w} \qquad \text{(equation 5)}$$

where μ_w is the linear attenuation coefficient of water and K_1 is a constant.[6] This relationship states that for homogenous substances, differences in μ from μ_w (the approximate linear attenuation coefficient of water at E = 73 kev, is $190 cm^{-1}$) are linearly amplified by a factor of K_1, in order to establish the CT number, which can then be displayed as some relative light intensity on a CRT screen. In the original EMI numbering scale, (Air = -500, Water = 0, and Bone = +500), K_1 is approximatley 2630 and K_1/μ_w is approximately 500. The exact relationship between the CT number and the linear attenuation coefficient is complicated by the fact that the CT system uses a polychromatic spectrum of x-ray energies (for example, an x-ray tube operating at 100 kVp, 120 kVp, or 140 kVp). Because the x-ray photons in this energy spectrum are subject to both photo-electric ($\mu_{\rho\varepsilon}$) and compton (μ_c) interactions, the total attenuation coefficient will be given by:

$$\mu = \mu_{\rho\varepsilon} + \mu_c \qquad \text{(equation 6)}$$

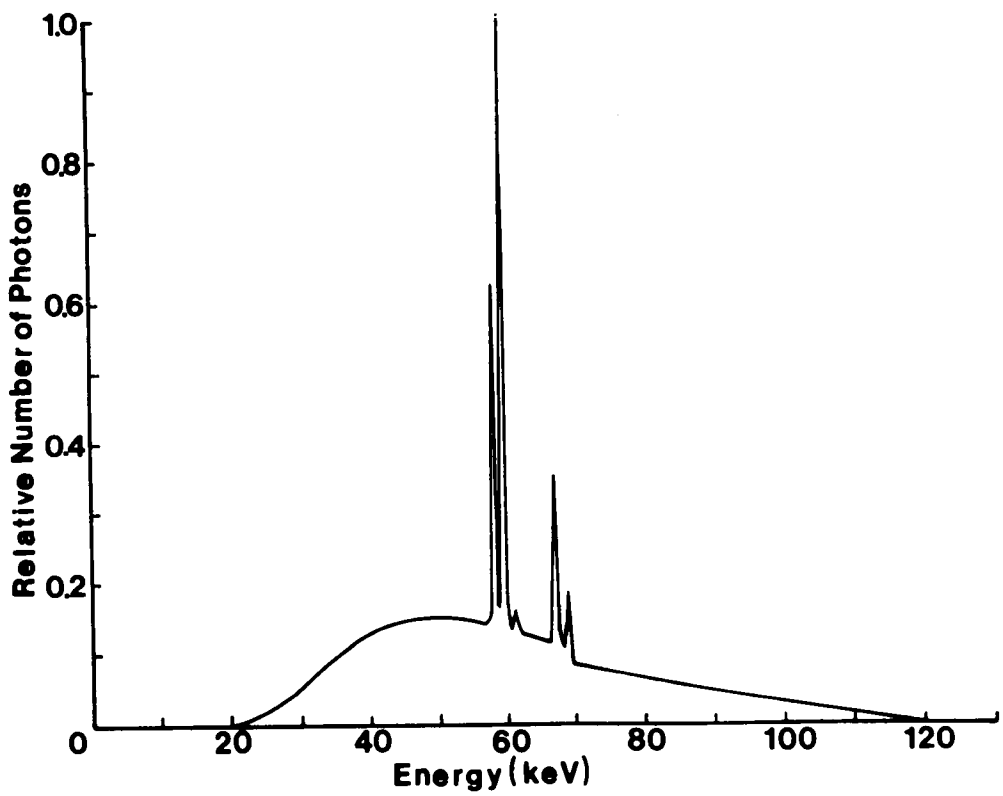


Fig. 5.

Diagnostic x-ray spectrum at 120 kVp with a half-value layer of 6.7 mm Al.[7]

Then, since $\mu_{\rho\varepsilon}$ is strongly dependent on atomic number (e.g. Z^3) as well as electron density (N_ε) and μ_c is primarily dependent on electron density (N_ε), the determined values of CT number (i.e. μ) will represent a complicated dependence on both the atomic number and the electron density of the material attenuating the x-ray spectrum. This subject has been discussed in an article where the EMI values were examined for correlation with physical density and atomic number cubed (Z^3).* 8

Another question of the applicability of this linear relationship is that the x-ray intensity distribution may vary over the object volume, and therefore one needs to know the relative x-ray intensity distribution at each point within the object volume before one can predict the (weighted average) attenuation coefficient that will be assigned to the particular volume element. An example of this position dependent effect can be seen in Figure 8 by the differences in the images of the same nominal beads, which differ only in their depth within their own relative volume elements. *9

In order to try to determine the contrast reaching the eye between the signal region and the background region in Computed Tomography, one can use the general CT relationship given by equation (5) and assume that the luminance display function is, to a first approximation, given by a linear function. One finds that the contrast is given by the difference between signal (μ_S) and background (μ_B) attenuation coefficients divided by the difference between the background and the lower window setting attenuation coefficient (μ_L). (See Appendix II)

Fig. 6a.

Radiograph of 3M skull phantom where no attempt has been made to reduce scattered radiation.

$$C_{CT} = \frac{\mu_s - \mu_B}{\mu_B - \mu_L} \qquad \text{(equation 7)}$$

The term in the denominator of this CT equation (μ_B minus μ_L) is an extremely important parameter in CT imaging because, whereas the normal contrast latitude loop of a radiographic screen film is visually invariable after photographic development; within a CT system, the contrast-latitude loop is left variable and subject to operator selection. Thus, in choosing to examine some given subregion of the continuum of attenuation values found between air and bone, one is basically selecting a "window" or subset of values to view at a given time (for example, a high contrast and low latitude; or a low contrast and a wide latitude view of the same object). There seems to be no question that this freeing of the viewer from a given contrast-latitude range is an important attribute of CT systems.

Fig. 6b.
Radiograph of 3M skull phantom where a 12:1 x-ray grid is used to remove scattered radiation.

Fig. 7.
Radiograph of 3M skull phantom with x-ray beam collimated to a small area around the sella turcica. Note the increased contrast as compared to Fig. 6a.

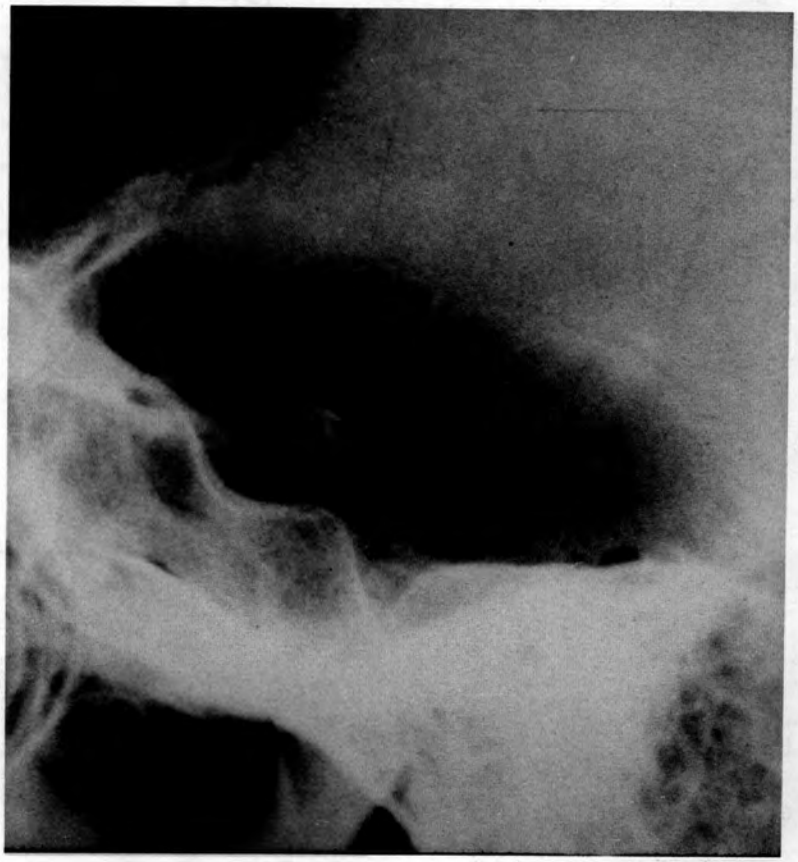


Fig. 8.
CT scan of beads at various depths with scan field showing positional (Z) dependence.

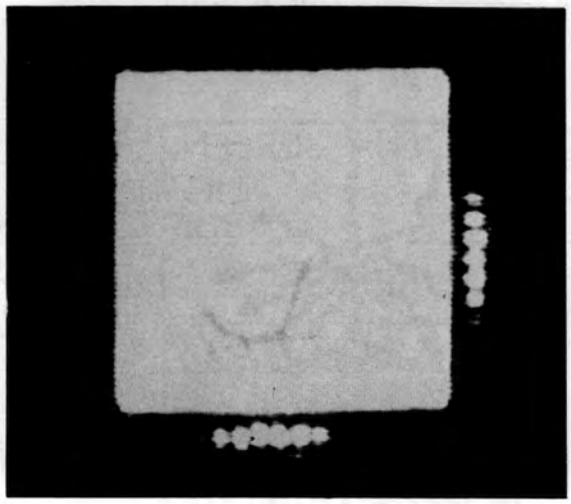


Fig. 9.
Signal size necessary for equal contrast in radiography and CT.

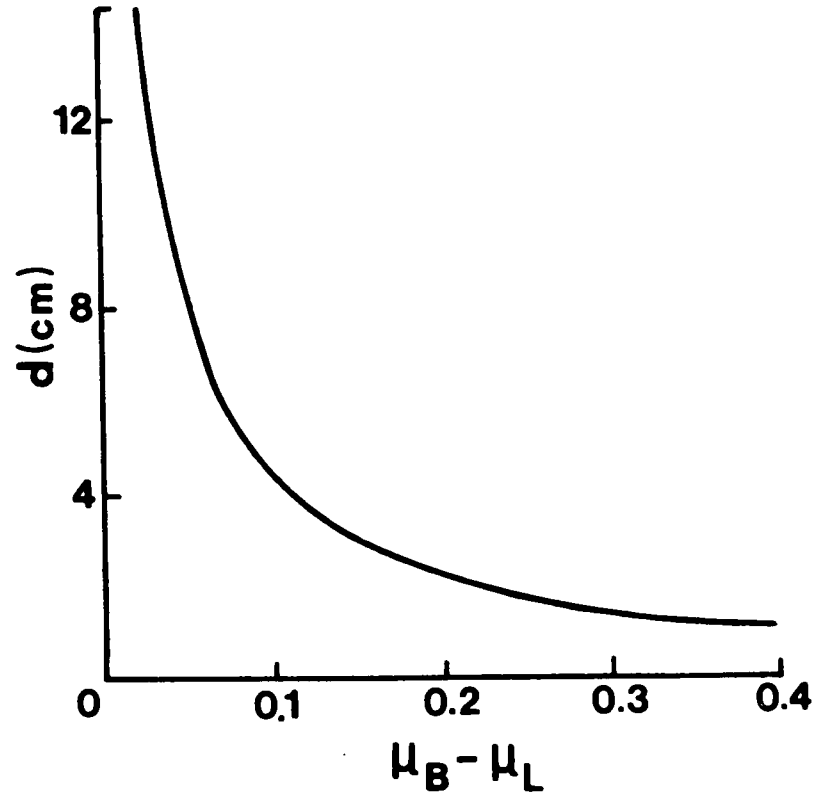
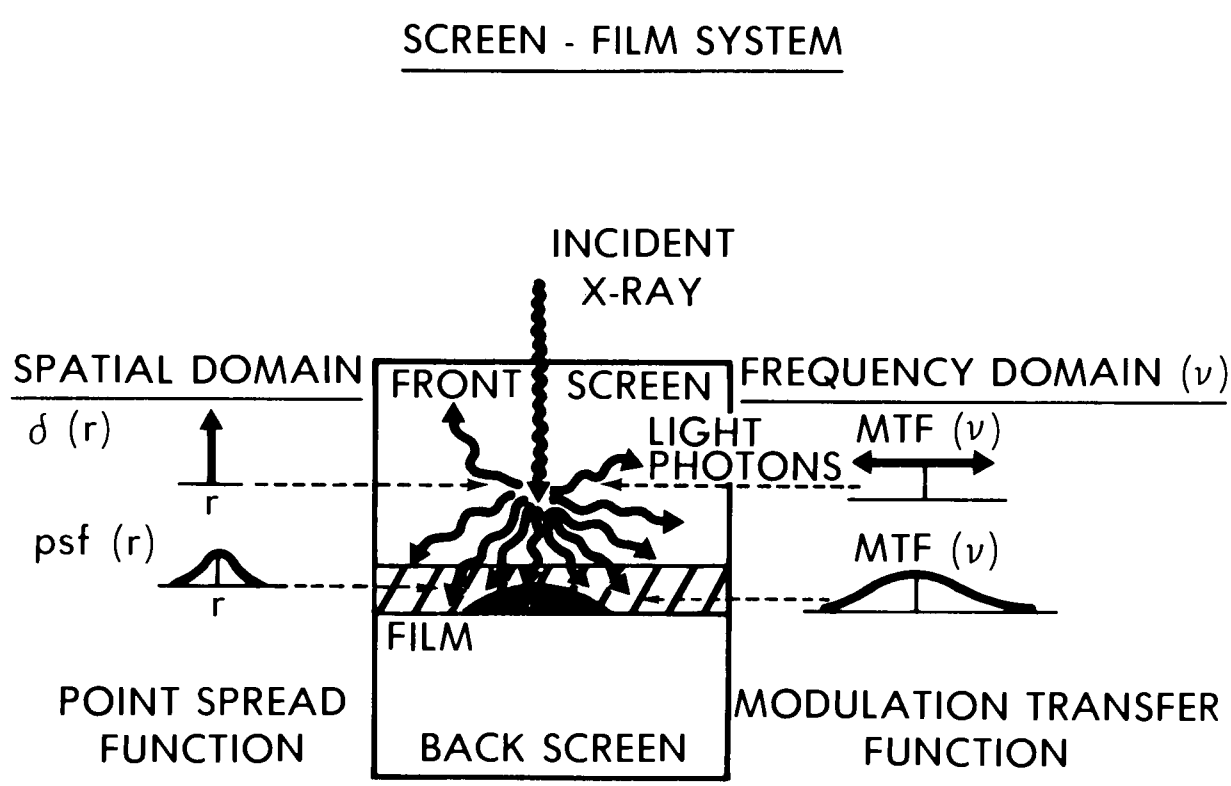


Fig. 10.
Schematic illustration of conversion of x-ray energy to light photons and accompanying loss of resolution.



Comparison of Contrast in Computed Tomography and Radiography

If one restricts attention to signals much larger than the resolution degradation function, and one now uses the relationships between the visual contrast given by radiography, divided by $G(\mu_B - \mu_L)d$, thus:

$$C_{CT} = C_R / G(\mu_B - \mu_L)d \qquad \text{(equation 8)}$$

This relationship (Fig. 9) predicts that contrast in Computed Tomography is greater than that in Radiography whenever the product of G, $(\mu_B - \mu_L)$, and d is less than unity. Then, in the sense that a typical window setting, μ_B is typically on the order of .02 cm^{-1}, this equation shows that in terms of visual contrast, radiographic signals will have to have extremely large diameters (several centimeters or larger), and/or have high differential attenuation coefficients in order to present a visual contrast to compete with Computed Tomography. In fact, it is probable that this high contrast property of CT systems-enabling demonstration of very small differential changes in the attenuation coefficient of tissue-is one way in which CT systems have made a significant breakthrough in the diagnostic arena. On the other hand, let us note that the parameter $(\mu_B - \mu_L)$ involved in the so-called window settings (choice of contrast and latitude in the CT system) is obviously being used to increase the relative performance of CT versus radiography. If one were to use other means of expanding the contrast scale in radiography, such as using certain kinds of television contrast stretching techniques, one might then, under certain conditions, hope to decrease the contrast advantage of CT. One may then ask what limits the ability to do this? The answer is found in the fact that any image processing technique including those used in Computed Tomography and those that one might seek to apply to radiography, will be limited by the signal to noise ratio levels that are an intrinsic property of the acquired information. Thus, if the noise level is very large then it probably makes little sense to augment or stretch the contrast of objects within the image, because, one will be essentially increasing or blowing up the noise fluctuations as well as the signal contrast. One must then also ask, does CT have inherent advantage over radiography in terms of signal-noise ratio?

NOISE IN RADIOGRAPHY AND COMPUTED TOMOGRAPHY

Noise in Radiography

The blurring or the unsharpness which accompanies the transducing operation of a radiographic screen/film combination, shown schematically in Fig. 10, also introduces changes in the appearance and amplitude of the statistical x-ray fluctuations which constitute the primary source of "noise" in the radiographic x-ray beam. In a sense, these fluctuations are also "imaged" by the screen. Sharply varying (uncorrelated point to point) noise fluctuations in the image are not found; rather, the noise (or density fluctuations) take on a characteristic mottled (or correlated) appearance. Instead of "salt and pepper" sharply varying density values across a uniformly exposed film, relatively larger areas of "Quantum mottle"-high and low density patches are found. In many ways, the blurring operation can be considered as a "noise smoothing" effect.

As mentioned previously, there are certain fundamental limitations imposed upon image quality by the constraint of having to use a finite radiation dose

in most radiographic procedures. A finite radiation dose implies an upper limit to the number of x-ray photons that will be stopped by the recording system (the image receptor). As is well-known, in most photon-limited imaging systems, a basic uncertainty in information transfer results from the Poisson statistics characterizing the x-ray beam. A mean number of absorbed counts per unit area, N, results in an expected standard deviation (s) in the number of counts around that mean value, given by the square root of N, thus: $s = (N)^{\frac{1}{2}}$.

In terms of the light transmission (T) reaching the eye, the relative (noise) fluctuations around the mean transmission value ($\sigma(T)/T$) are given by the following equation: * 10

$$\frac{\sigma(T)}{T} \simeq \frac{G}{n_x^{-1/2}} F \qquad \text{(equation 9)}$$

It can be noted that, as in the equation describing radiographic contrast, the parameter G, and a parameter related to the size of the signal (in this case the area of the signal being searched "a") are found. In addition, the expected number of absorbed photons per unit area $\bar{n}_x$, as well as an interesting parameter "F" which is related to the inherent spatial frequency transfer capability tends to transfer the Poisson noise fluctuations a little better because such noise has a significant high spatial frequency component. In addition, though, the "F" factor introduces a complication to the evaluation of noise in radiographic imaging. That is, the modulation of the noise statistics by the MTF gives the characteristic shape, or appearance to the noise fluctuations, often described by the term, "Quantum Mottle".* 11 "Quantum Mottle" is important because the preception to "noise" is influenced not only by the amplitude or excursions of the noise around the mean value; but also, by the characteristic mottled appearance or characteristic pattern of the noise. Certain noise patterms seem to be more objectionable in certain viewing tasks than other types of noise.

The use of faster screens or faster film to reduce the effect of patient motion and/or to allow the use of lower kVp while keeping the imaging task within the finite generator output of the x-ray system, is often advocated. As a first approximation, for a given film density, and for a given phosphor type, the fewer the number of x-ray photons absorbed in the screen, the more the noise amplitude will tend to increase. In considering "white noise", that is noise which varies sharply from point to point, it can be assumed that for small signals, increased noise amplitude reduces certain image detail at about the same rate as a corresponding increase in contrast, enhances image detail.* 12 This concept is usually considered as the "signal-to-noise ratio" approach. In this approach, it is assumed that a figure of merit of a system is the relative ratio of the contrast of the signal divided by the relative noise amplitude of the surrounding (competing) background. If it is assumed that the noise amplitude in an image will increase as the square root of the speed, then for every increase in speed by a factor of 4, twice as much contrast in the signal will be needed to detect a given signal as well on the 4X system as on the image obtained from the conventional speed system. Although this approach is overly simplistic, it does correctly point out that increasing contrast may result in increased noise. As the signal size becomes increasingly small, and looks more and more like the typical noise excursions, then the preceeding signal to noise ratio approach should become more and more exact. It should be noted that in the case of large signals, on the order of 1 cm. or larger, the real influence of the finite photon flux, probably becomes of smaller and smaller consequence. In particular, note that the relative noise fluctuations for 1 cm. diameter signals for present medium speed film/medium speed screen combinations are less than one part in a thousand, (this results from stopping about 10^7 x-ray photons per cm^2 of screen). However, the other problem with noise is that it is well

known that the human eye is lookihg essentially for borders or edges (high spatial frequencies) and thus, even a large signal having a low contrast unsharp edge may have important edge information obscured by noise fluctuations.

Faster screens for radiography may be sought in essentially three major ways: one, by using a thicker screen; two, by using a phosphor with higher stopping power; and three by using a phosphor with a higher conversion efficiency of x-ray energy to light energy. In the first method of using a thicker screen, the increased spreading of the light in reaching the film can cause decreased resolution; that is, an increased blurring effect. Whereas, in the second method of utilizing a higher stopping power, a thin screen may perhaps, maintain the optical resolution properties of the screen and still absorb the same or more x-ray photons. The third method, utilizing a phosphor with a higher conversion efficiency, has the effect of requiring fewer x-ray photons to obtain the same light output and therefore results in increasing the noise level. The speed of the system might also be increased by increasing the sensitivity of the film so that for a given light output from the screen, one has a higher latent image production in the film emulsion, an approach which also produces statistical effects on the noise levels.

Figure 11 is used to try to illustrate the complicated inter-relationship between resolution and noise in radiographic imaging systems.

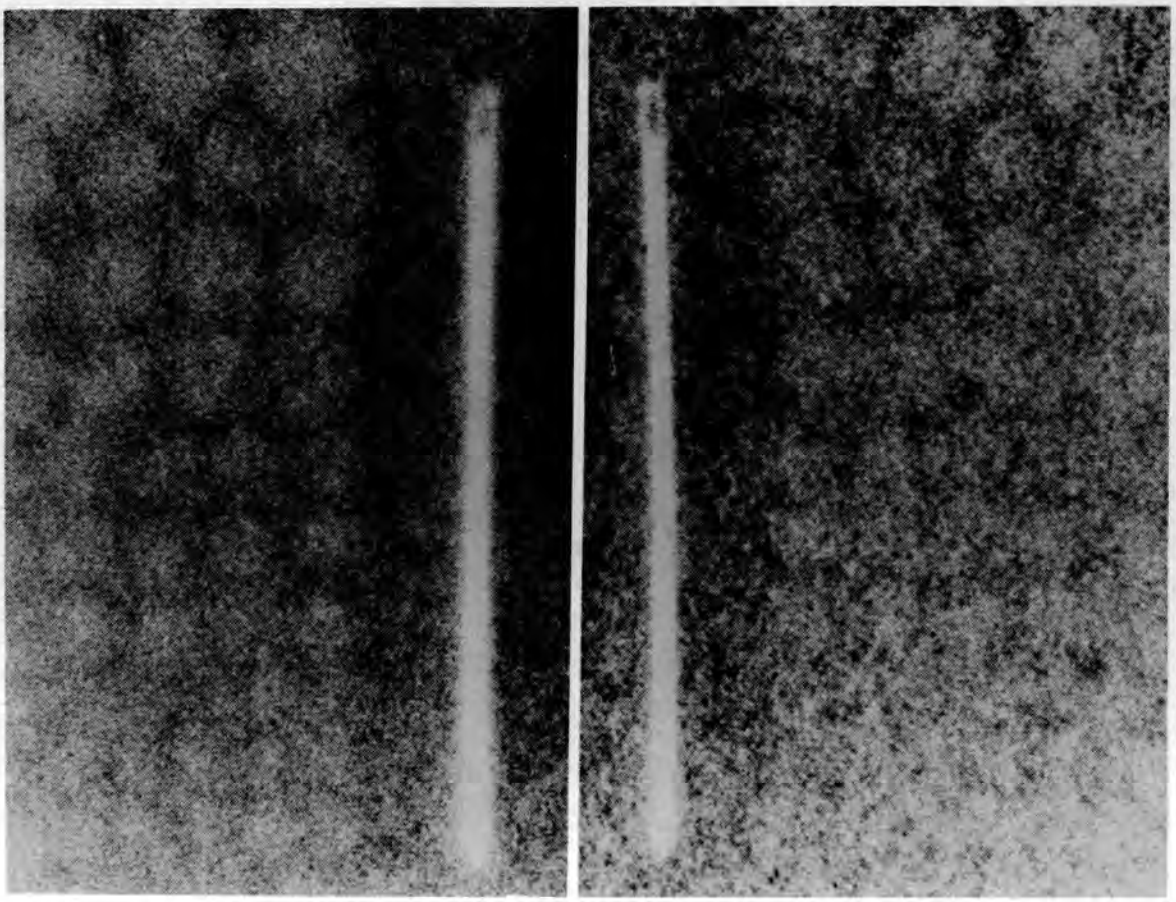


Fig. 11.
Radiographs of a steel needle and plastic beads taken with medium speed screens and fast film (right) and fast screens with normal speed film (left). The needle is seen best on the high resolution on the high resolution system.

The radiographs shown in Figure 11 are obtained by using two different film and screen combinations that have relative film and screen speeds which are matched so that the overall system speed (and thus motion unsharpness) are equal. Both screens are calcium tungstate intensifying screens. It is found that the image of contrast filled vessel-type objects (needles) is essentially resolution limited. This type of object is imaged better with the system with the higher resolution properties, that is a sharper system. The accompanying low contrast tumor-type objects (plastic spheres) which are essentially noise limited are actually detected better when radiographed on the system which shown poorer sharpness and more smoothly varying noise. In addition, Figure 12 shows the corresponding image of a single low contrast tymor-type object in an isolated position on each of the two systems. It may be seen that the system with a higher intrinsic sharpness does not always afford the best detection for low-contrast noise limited detail. This subject and its effect on human observer detection performance (ROC curves) has been discussed at length in other papers. *[2,10] It may be loosely said that for high contrast vessels where sharp border information is necessary that "resolution" is a key factor, whereas, for small,low contrast tumors with diffuse borders, reduced noise becomes imcreasingly more important even at the expense (and somewhat as a positive consequence) of less sharpness.

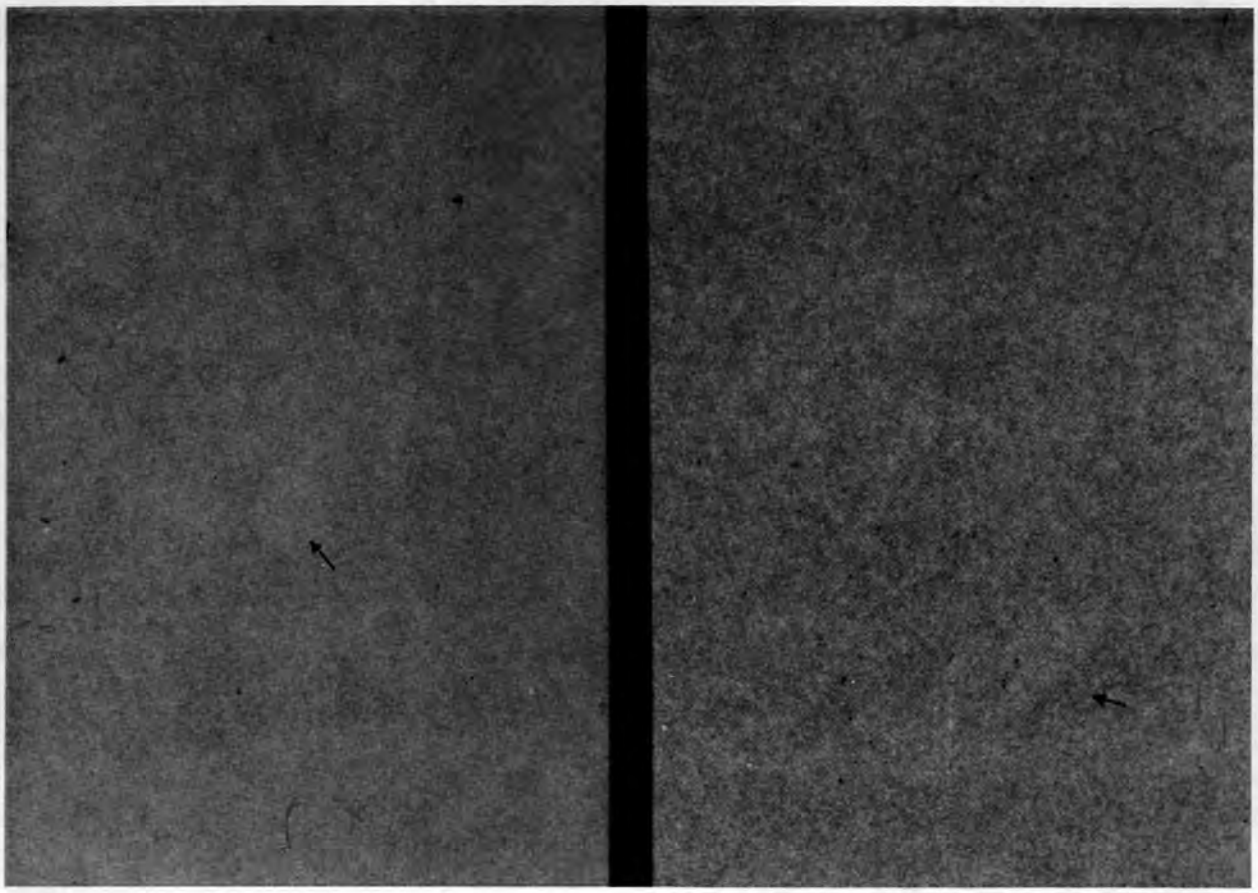


Fig. 12.
Radiographs of single beads by systems used for Figure 11, demonstrating the advantage of diminished noise for low contrast noise limited detail.

Noise in Computed Tomography Systems

An equation often used to describe the relationship between the standard deviation of the noise or fluctuations in attenuation values, $\sigma(\mu)$, in a CT scanner with matrix element dimension (w), slice thickness (h), and the patient dose (D), is given by:

$$\sigma(\mu) \propto \frac{1}{Dw^{3/2}h^{1/2}} \qquad \text{(equation 10)}$$

This relationship, however is overly simplistic in that it ignores the important question of the character of the noise that results from the combination of photon statistics and the choice of filter function. A rigorous treatment of the noise in a CT scan requires consideration of the auto-covariance properties of the noise (or the corresponding noise power spectrum).

It has been shown by Tanaka that a more realistic approach to the relationship discussed above is:

$$\sigma(\mu) \propto \frac{1}{D\ h^{\frac{1}{2}}\ (FWHM)^{3/2}} \qquad \text{(equation 11)}$$

where, FWHM now represents the Full Width Half Maximum of the point spread function of the reconstructed image, rather than the matrix element width (w). The filter function will have a significant effect on the amplitude and character of the noise component in a CT scan. * [13]

If we again assume that the light intensity, (T), reaching the eye from a region of the CT image is approximately linearly related to the CT value (§), and therefore to the linear attenuation coefficient (μ), then it can be shown that for signal area "a" which is relatively large compared to the pixel area, w^2, that the expected standard deviation of μ - values in area "a", $\sigma(\mu,a)$, will be given by:

$$\sigma(\mu,a) = \frac{w}{a^{\frac{1}{2}}}\sigma(\mu) = \frac{K_2}{(D\ wha)^{\frac{1}{2}}} \qquad \text{(equation 12)}$$

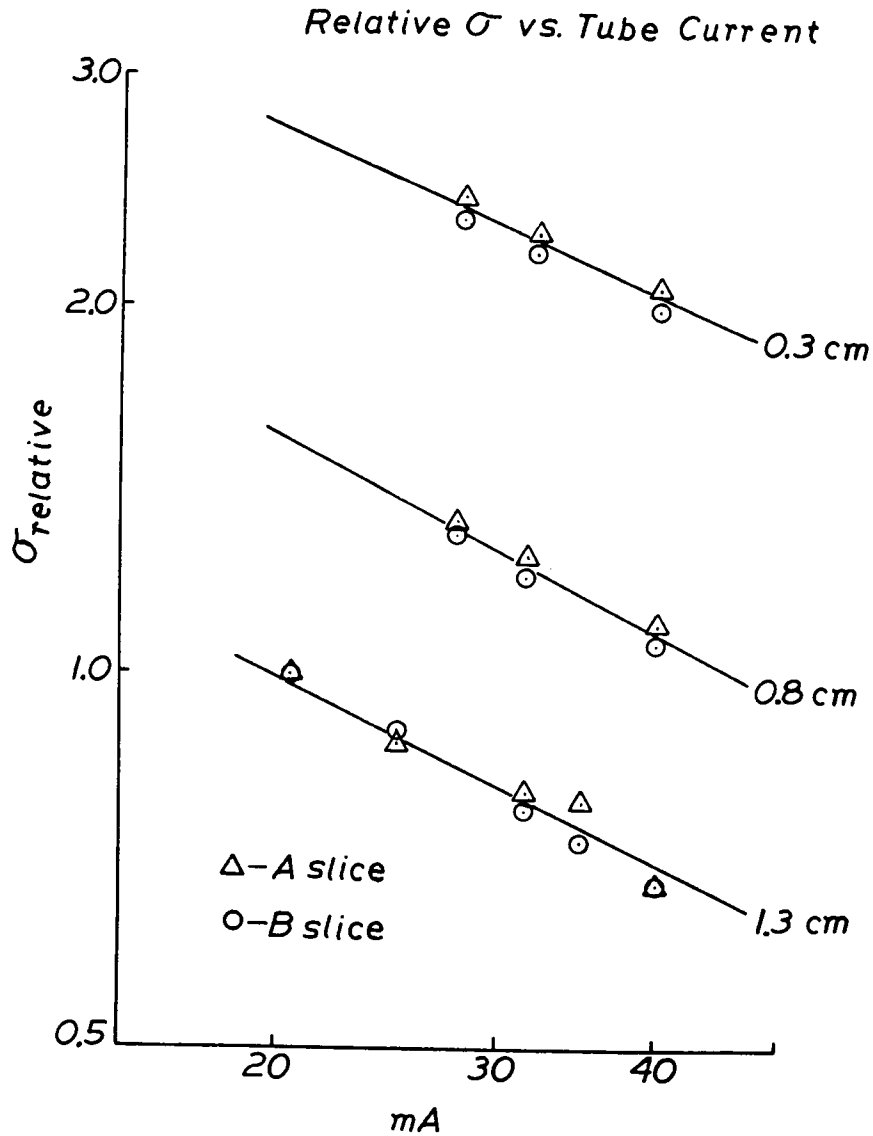
Some experimental data showing the relationship between noise amplitude (σ), slice width, and tube current ($\propto$ to patient dose) is shown in Fig. 13. If one considers noise in current CT scanners, then for D = dose in rads, w = matrix dimension width in cms, h = the slice thickness in cms, and, a = signal area in cm^2, and estimate K_2 on the basis of an EMI Mark I head scanner as approximately 10^{-4}; then, because $T \simeq K_3\ (\S - \S_L) \simeq K_4\ (\mu - \mu_L)$ (equation 13)

$$\frac{\sigma\ (T)}{T} \doteq \frac{\sigma(\mu)}{\mu - \mu_L} \qquad \text{(equation 14)}$$

and in terms of relative fluctuations, around the mean background, light intensity, T_B,

$$\frac{\sigma\ \ (T)}{T_B} \doteq \frac{10^{-4}\ (D\ wha)^{-\frac{1}{2}}}{\mu_B - \mu_L} \qquad \text{(equation 15)}$$

Fig. 13: Dependence of relative σ on tube current (mA) and mean slice thickness (Δz) for the EMI Mark I head scanner.



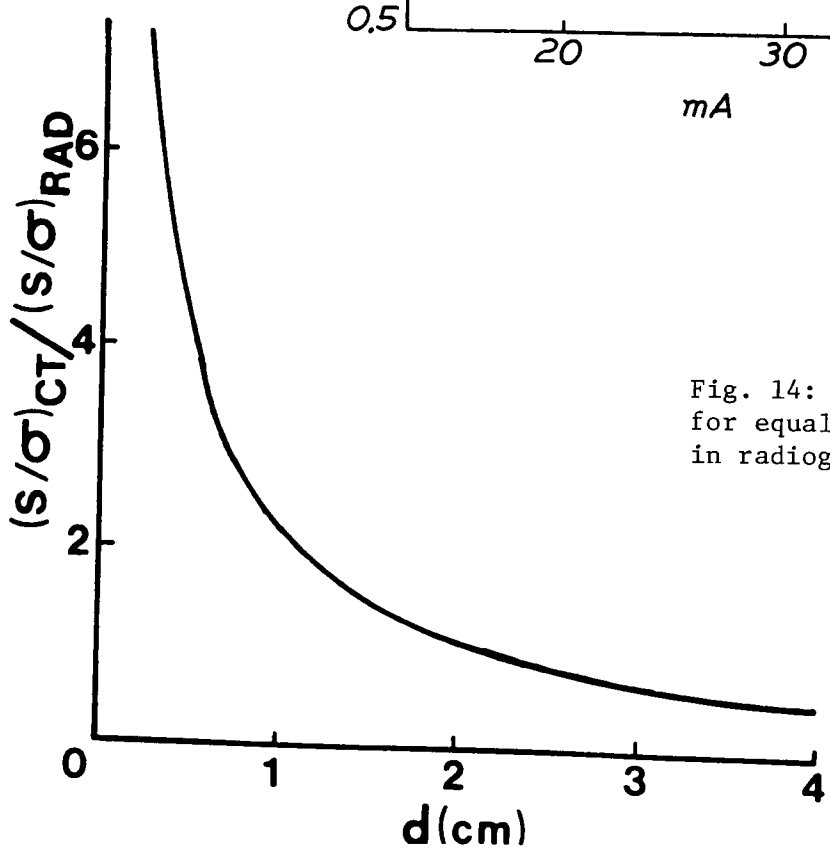


Fig. 14: Signal size necessary for equal signal to noise ratio in radiography and CT.

Let the signal to noise ratio, d', be defined as the ratio of signal contrast to noise amplitude, then,

$$d'_{CT} = 10^4 \ (\mu_S - \mu_B) \ a^{\frac{1}{2}} \ (whD)^{\frac{1}{2}} \qquad \text{(equation 16)}$$

$$d'_R = (\mu_S - \mu_B) d \ \bar{n}_x^{\frac{1}{2}} \ a^{\frac{1}{2}} \qquad \text{(equation 17)}$$

Then for the same x-ray spectrum, if we assume, $\bar{n}_x^{\frac{1}{2}} = 3 \times 10^3 \ cm^{-1}$, w = .15 cm, h = 1.3 cm, D = 2 rad and in the absence of a significant scatter fraction;

$$\frac{d'_{CT}}{d'_R} = \frac{10^4 \ (whD)^{\frac{1}{2}}}{d \ \bar{n}_x^{\frac{1}{2}}} \simeq \frac{3}{d} \ (whD)^{\frac{1}{2}} \simeq \frac{2}{d} \qquad \text{(equation 18)}$$

This relationship is shown in functional form in Fig. 14. In the presence of scatter, this relationship must be modified to:

$$\frac{d'_{CT}}{d'_R} \simeq \frac{2}{d} \ \frac{(E_B + E_\phi}{E_B} \qquad \text{(equation 20)}$$

The exact relationship between the signal to noise ratios is obviously dependent on the degree of scattered radiation, and the choice of values for K_2, w, h, d, D and $\bar{n}_x$; the inverse relationship of this ratio on the signal size seems clear. Thus, at some signal size (on the order of a few centimeters) the signal to noise ratio of Computed Tomography should become less than that of radiography. The principle signal to noise ratio advantage of Computed Tomography would appear to be found in enabling detection of a signal with a diameter on the order of 1 cm or so. Several caveats must be kept in mind in this analysis: (1) even if the signal to noise ratios were equal for Computed Tomography and Radiography, one might still have to use contrast enhancing techniques in radiography to make the signal contrast perceptible - however, the basic signal-to-noise <u>would</u> be adequate to do so; (2) The degrading effect of the scatter contribution is obviously very important in radiography and its magnitude could even require signals to be an order of magnitude larger than previously suggested before radiography could compete (for example in dense regions of the abdomen) with Computed Tomography; (3) The analysis so far has ignored potential differences in spatial resolution and motion unsharpness; (4) the simple model does not present differences in the perceptual tasks of "projection" radiography versus "tomographic" imaging.

RESOLUTION FACTORS IN RADIOGRAPHY AND COMPUTED TOMOGRAPHY

Resolution in Radiography

It was mentioned in the previous discussion on contrast that one of the problems in trying to obtain increased subject contrast by lowering kVp is the increased motion unsharpness which results from the increased time which must be used to generate adequate x-ray penetration of the patient There are known to be three basic sources of resolution degradation or unsharpness in screen/film radiography: (1) Motion unsharpness - the blurring of a point due

to the motion of the patient during the exposure time of the procedure; (2) Receptor unsharpness - the inherent blurring operation caused by the transducing or conversion of x-rays into light within the intensifying screen; and (3) Geometric unsharpness (focal spot unsharpness) - the blurring due to the finite size and x-ray intensity distribution of the x-ray focal spot. Combinations of the above sources generally lead to an overall blurring or defocusing effect.

This blurring or unsharpness may be characterized by either its spatial extent, which is generally the blur size or "point spread function", or the same information may be characterized in the spatial frequency domain by the Modulation Transfer Function (MTF) which describes how the spatial frequencies which make up an object are transferred by the image recording system.[14] Thus, for example, high spatial frequency components which generally make up the important edge structures in radiography are generally poorly relayed through the radiographic system. Typical MTF data for screen/film systems is shown in Figure 15.

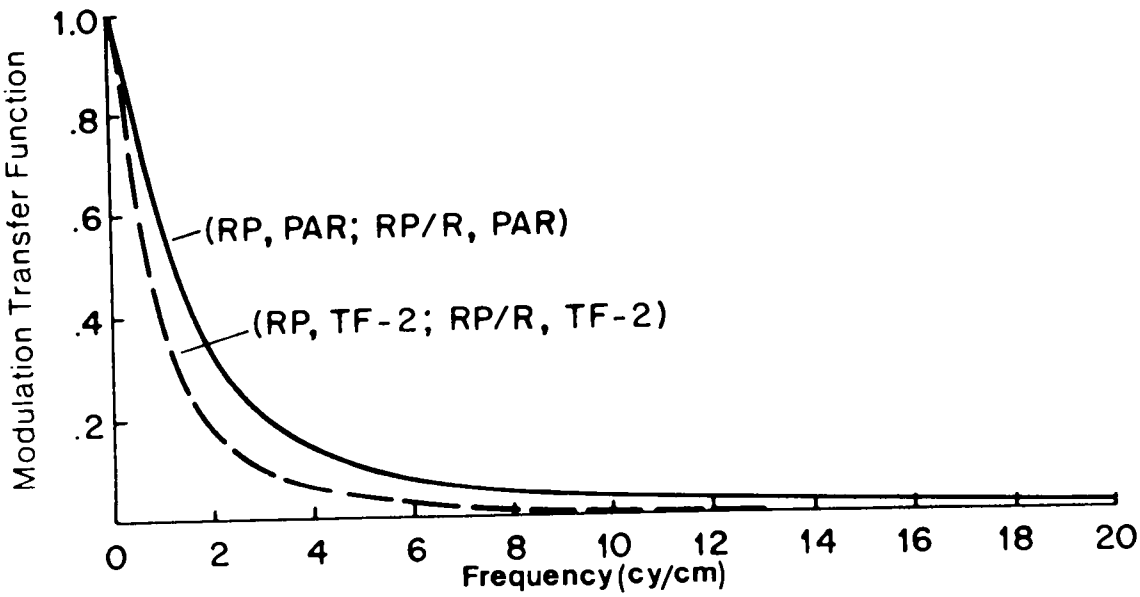


Fig. 15: Modulation Transfer Functions for medium speed screens (Par) and high speed screens (TF-2).

The schematic diagram of Figure 10 can again be examined to see how the incident x-ray photons are transduced or converted in an intensifying screen to a burst of light photons which then produce the latent image within the film. The transducing operation is itself a blurring operation, because each incident x-ray photon confined essentially to a point dimension causes a burst of light photons, which before reaching the film, spread over considerable spatial area and therefore, degrade the information concerning the exact spatial origin of the x-ray photon. Instead of a sharp, one-to-one correspondence between the incident x-ray and the developed image point, we have a

"point source" imaged as a blur or a "point spread function" on the films as indicated schematically in the figure. It turns out that to a first approximation the overall system point spread function might be considered a Gaussian-shaped function. This approximation is convenient in that a rotationally symmetric Gaussian distribution may be fully described by a single parameter, namely, its standard deviation (σ_R) which is a measure of the width of the distribution and the extent of the blurring. A specification of σ_R for each of the three blurring sources mentioned previously, may be used to describe the relative blurring effect of each source of unsharpness; σ_m, σ_r & σ_g. In addition, by assuming that the individual blur functions are Gaussian, then the overall unsharpness or blurring function is given by the root mean square (rms) of the individual blur contributions.

$$\sigma_R = \sqrt{\sigma_m^2 + \sigma_r^2 + \sigma_g^2} \qquad \text{(equation 21)}$$

Table 2 indicates the way in which the various blur sources mentioned previously combine into an overall general blur or degradation function for radiography.[15]

TABLE II

$$\sigma_R = \left(\frac{\sigma_r^2}{M^2} + \sigma_m^2 t^2 + \sigma_g^2 \left(\frac{M-1}{M}\right)^2 \right)^{\frac{1}{2}} \qquad \text{(equation 22)}$$

Magnification	Relative Exposure Time	Focal Spot Unsharpness		Screen/film Unsharpness	Motion Unsharpness	Total Unsharpness
M	t	σ_g		σ_r	σ_m	σ_R
1.0	0.2 secs	1 mm	Slow Screen	0.33 mm	5 mm/sec	1.053 mm
1.5	0.3	1		0.33	5	1.552
2.0	0.4	1		0.33	5	2.068
1.0	0.2	2		0.33	5	1.053
1.5	0.3	2		0.33	5	1.656
2.0	0.4	2		0.33	5	2.242
1.0	0.1	1	Fast Screen	0.66	5	0.828
1.5	0.15	1		0.66	5	0.931
2.0	0.2	1		0.66	5	1.166
1.0	0.1	2		0.66	5	0.828
1.5	0.15	2		0.66	5	1.096
2.0	0.2	2		0.66	5	1.452

In particular, a motion unsharpness value of σ_m = 5mm/sec is used. Certain advantages can be obtained from magnification studies, in that the focal spot, image receptor and patient motion blur factors may be carefully selected so as to produce some reduced overall unsharpness. Moreover, magnification is often used with an airgap technique which reduces the scatter contribution without the need for grids. Of course, it can be noted from Table 2 that the x-ray generator output and the focal spot size (both the nominal and actual size) will be important in determining the limits of magnification, as discussed in other papers.[15]

Resolution in Computed Tomography

In discussing the spatial resolution of computed tomography systems, it should be kept in mind that CT systems are neither strictly linear nor isoplanatic (output is not independent of position).

Keeping in mind the caveats concerning isoplanacity (isovolumetricity) and linearity, we may consider the spatial resolution of the CT system by using the following resolution test objects mounted in the "Resolution Module" (Figure 16) a phantom designed by the authors.[16]

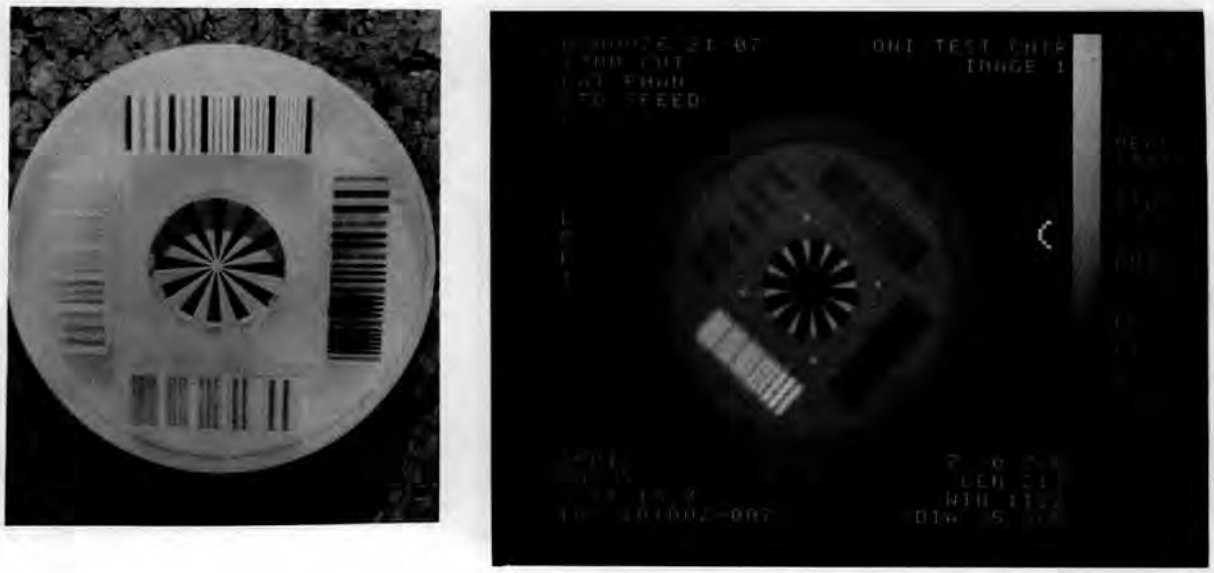


Fig. 16: The "Resolution Module"
(a) Photograph (b) CT Scan

An indication of the spatial frequency transfer capability of a CT system can be obtained by taking the Fourier Transform of the average data resulting from several CT scans of a thin wire object aligned with the vertical axis (z) of a centrally located volume element. The wire object then constitutes a point source input in the axial scan (x,y) plane of the CT system. Care must be taken so that the projections through the voxel containing the point source (wire object) do not exceed the dynamic range of the CT scanner. Furthermore, the point source should be considerably smaller than the intrinsic resolution of the system. For these reasons, for pixels on the order of 1 mm x 1mm, one needs to restrict the wire diameter to about 300 microns or less.

Thus, in the resolution section, we find a set of four point source objects (wires) for determination of point source response functions and corresponding Fourier Transforms (MTF), as illustrated in Figure 17.

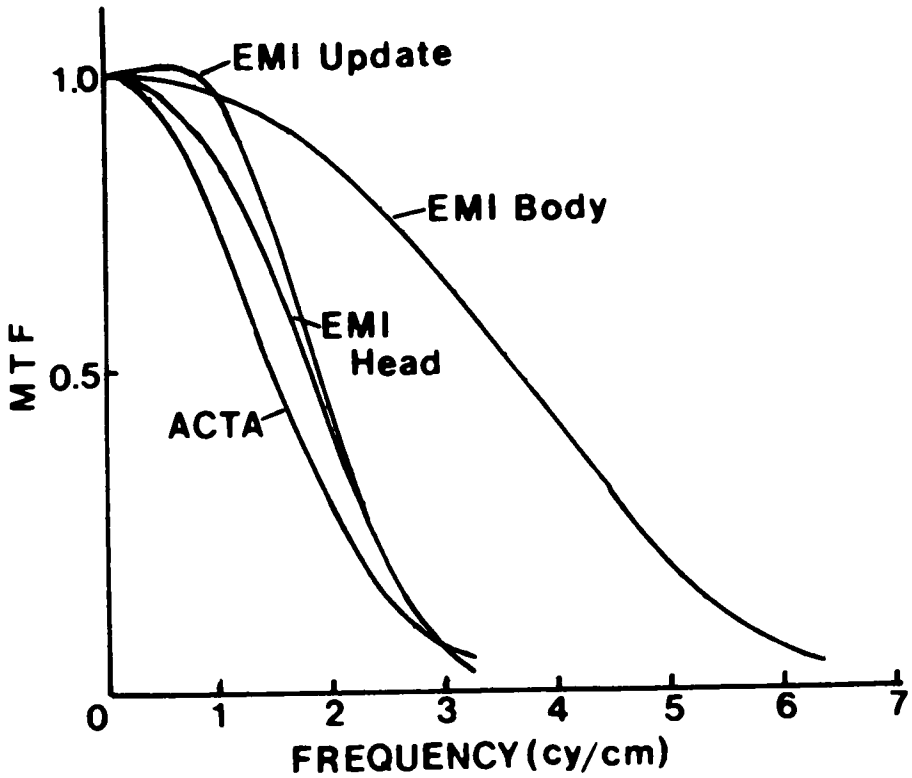
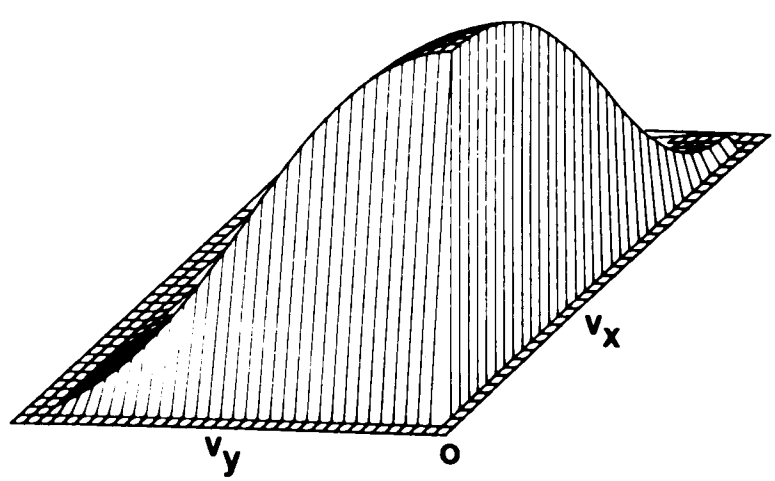


Fig. 17: Fourier transform of point spread function data determined from scanning thin wire objects in the indicated CT systems.

For reference purposes, the two-dimensional MTF describing the transfer characteristics found with the EMI Mark I head system operating at 120 kVp and using a 160 x 160 scanning matrix (that is a 1.5 mm x 1.5 mm pixel size), as appears in Figure 18 shows at least a 10% modulation transfer of frequencies below three cycles/cm.

Fig. 18: Two dimensional MTF of EMI Mark I head scanner. Each grid line increment corresponds to 0.104 cycles/cm.



As stated previously, because of questions of linearity and isoplanacity, such MTF data may be dependent on the relative spatial position of the point source and will not necessarily be generally useful in predicting output data from known input data (unless the positional dependence and volumetric averaging are taken into account).

Moreover, it should be noted that because CT systems generally use a square matrix element, the point spread function and the MTF will be asymmetric and therefore attempts to describe the MTF of the system by a one-dimensional Fourier transform of a line spread function will be dependent on the relative orientation of the line source within the scan plane and therefore not generally valid, unless the line spread function and corresponding Fourier transform is obtained for each possible orientation of the line source.

Within the "Resolution" module, one also finds two other test patterns of spatial resolution more amenable to on-line visual determination of spatial resolution cutoff frequency; these test patterns are, respectively, a series of resolution test gauges and a star pattern. The cutoff frequency of an imaging system is a useful but somewhat limited description of spatial resolution of imaging systems. In particular, the determination of cutoff frequency will depend on the intrinsic contrast level of the test pattern. The test is contrast dependent in that the particular cutoff frequency must have a contrast level above the visual threshold. The MTF measurement is a more comprehensive discriptor in that it offers information concerning the contrast (or modulation) transfer capability of a system at each spatial frequency of interest. Because of this contrast dependency, the resolution section contains resolution gauges of four different intrinsic contrast levels, namely: high, medium-high, medium-low, and low.

These resolution gauges provide an estimate of the intrinsic CT contrast level and separation distance necessary to visually resolve two adjacent objects.

The following spatial frequency targets are provided in *each* gauge: 2, 3,4,5,7 lp/cm where 1 lp/cm represents 1 full cycle of a square wave pattern, the contrast characteristics of which can be related to the more formal sinusoidal frequency cycle as described by the MTF.

The third type of test of resolution provided in the phantom is a central star pattern containing radially varying spatial frequencies up to 10 lp/cm. This pattern is designed to provide a convenient quality assurance test of spatial resolution. The composition of the star provides a high contrast (-1000→+600) pattern for determination of the "star cutoff" frequency.

It is worthwhile to consider the relationship among the preceding measurements of resolution in terms of the importance of single number measurements of resolution, such as cutoff frequency vs more complete descriptors of spatial resolution such as MTF. We can see from Figures 19 and 20 that one system can have a higher cutoff spatial frequency while having decreased modulation transfer characteristics at lower spatial frequencies. However, in terms of the overall perception of contrast within the picture we need to consider also the human visual describing function (pseudo MTF).

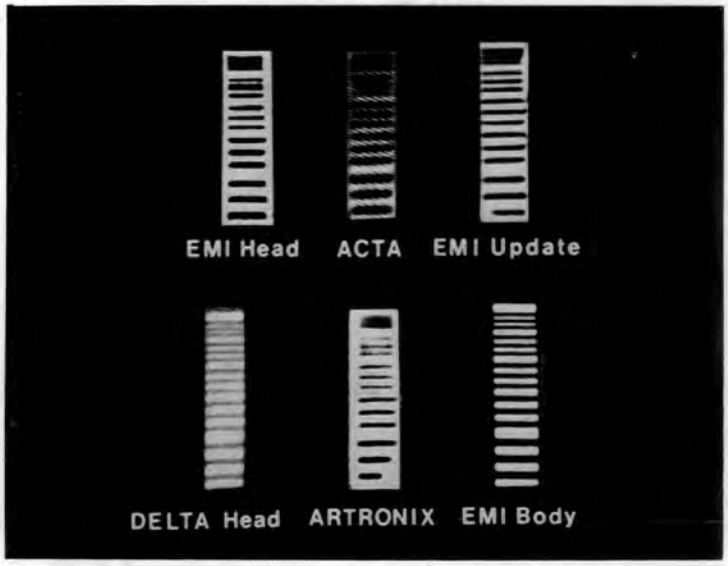


Fig. 19: Experimental cutoff frequency of several different CT scanners determined from an air-Lucite resolution gauge with frequencies of 1.25, 1.6, 2.2, 2.5 and 3.85 line pairs/cm.

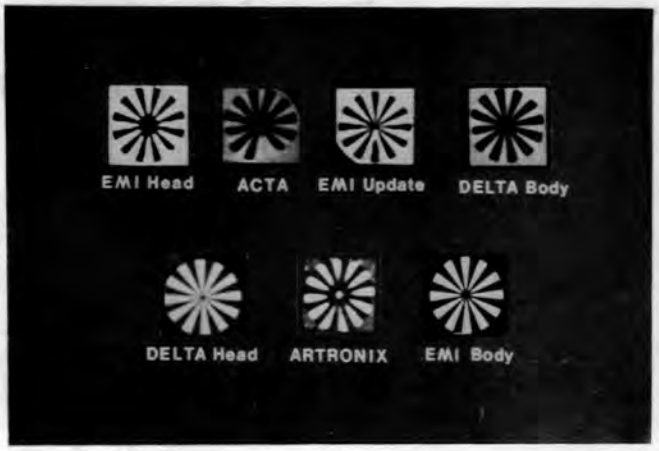


Fig. 20: Experimental cutoff frequency of several different CT scanners determined from prototype (air-Lucite) resolution star pattern. Note that a different star pattern is used for the Artronix scanner and several pictures have spokes missing due to the intrusion of water into the phantom.

In Figure 21, we can examine the difference in MTF between two systems that have different cutoff frequencies, then we can see that at normal viewing distances, the MTF of the human visual systems tends to match the region of enhanced MTF of the system with the lower cutoff frequency. In fact, when we apply an energy detection model to this data (i.e., the integral of the square of the image spatial frequency spectrum), we find that in certain cases the "perceived" small object contrast (such as an edge) could be enhanced significantly by the system with the lower cutoff frequency.

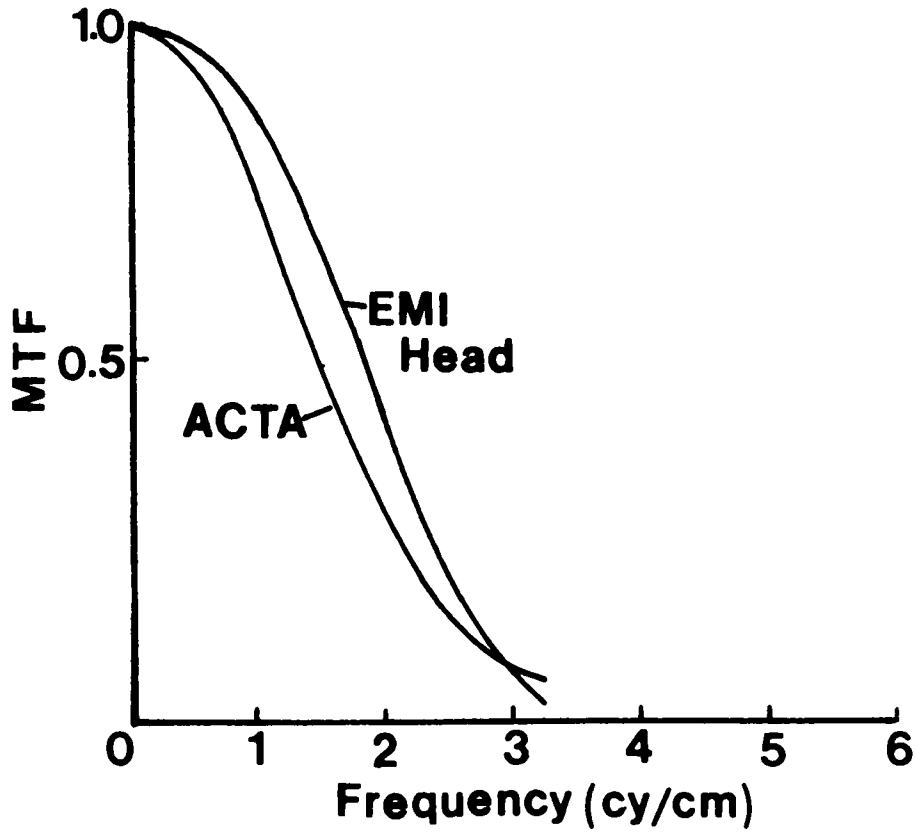
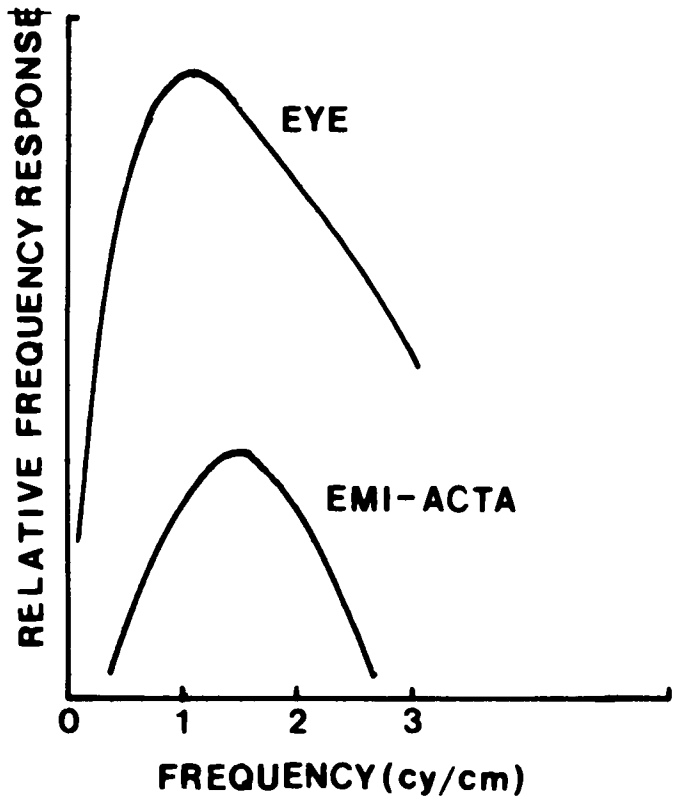


Fig. 21: Illustration of way in which differences in system MTF may be augmented by MTF of the eye. (a) above; MTF's of two different CT systems.

(b) right; Difference of MTF between the two systems along with MTF of the eye.



Limitations Imposed by Resolution in Radiography and Computed Tomography

From the preceeding discussion it can be noted that there is a real limit as to how small a signal one might presently hope to detect with CT systems because of both the volume averaging effect (the influence of slice width as well as the pixel) and the realization that spatial resolution of CT systems is presently much inferior to radiography. Thus for small objects where the contrast (in particular the edge perception) is strongly influenced by MTF, (i.e. related approximately to the integral of the square of the MTF), one finds a rather interesting effect of spatial resolution in that the contrast in CT and radiography will be weighted by the spatial resolution properties of the radiographic system compared to the spatial resolution properties of the CT system in a quadratic or square nature.

Then, assuming similar x-ray spectra, one has:

$$\frac{C_{CT}}{C_R} \doteq \frac{\sigma_R^2}{\sigma_{CT}^2}\left(\frac{1}{G(\mu_B-\mu_L)d}\right)\left(\frac{E_B+E_\phi}{E_B}\right)\left(\frac{1}{\forall}\right) \qquad \text{(equation 23)}$$

where, σ_R, σ_{CT} are measures of the relative width of the point spread function in radiography and computed tomography; and $\forall$ is a measure of reduction in contrast due to volume averaging in computed tomography.

To a first approximation, for signal area, and diameter, d, small compared to σ_R, σ_{CT}, w and h; one can estimate $\forall$ as:

$$\forall = \frac{6\ w^2h}{\pi d^3} = \frac{1.5\ w^2d}{ad} \qquad \text{(equation 24)}$$

and in the absence of scatter:

$$\frac{C_{CT}}{C_R} = .66\ \frac{\sigma_R^2}{\sigma_{CT}^2}\ \frac{a}{w^2h\ \ G(\mu_B-\mu_L)} \qquad \text{(equation 25)}$$

If one assumes G=2.5, w=.15cm, h=1.3cm, $\mu_B-\mu_L=.02\text{cm}^{-1}$ (i.e. a window of 100 on original EMI Mark I head scanner), $\sigma_{CT}=4\sigma_R$, then one can show that in the absence of scatter:

$$\frac{C_{CT}}{C_R} \doteq 28a \qquad \text{(equation 26)}$$

where a is in centimeters, and in the presence of scatter,

$$\frac{C_{CT}}{C_R} \doteq 28a\left(\frac{E_B}{E_B+E_\phi}\right) \qquad \text{(equation 27)}$$

Clearly, the calcified specks with diameters on the order of fractions of a millimeter will have superior contrast in radiography than in computed tomography.

Summary

This paper has attempted to show that the use of a relatively simple detection model and known physical image parameters of radiography and computed tomography predict that the intrinsic imaging advantage of computed tomography may be restricted to signals (tumors) on the order of several millimeters to a few centimeters. Several caveats must be kept in mind in attempting to extrapolate these results too far into clinical practice. These caveats include physical considerations such as: the degradation effects of scatter, motion unsharpness, longitudinal magnification, distortion and characteristic artifacts; as well as perceptual consideration such as the difference in pattern recognition tasks and structure noise arising in projection radiography versus (pseudo) slice images of computed tomography.

The results of the model do raise many important points that deserve future study and research. These important results include the dependence of relative image system performance on signal size - the fact that CT may find optimum use for signals on the order of 1 cm or so; whereas, radiography, especially when used with contrast enhancing and/or other image processing techniques, may be of advantage with relatively smaller and larger signals than those sizes with which CT seems to have an intrinsic signal to noise advantage.

Then too, one can see that typical approximate radiation exposures of 70-100 mR/film for the skull,and 500 mR/film for the pelvis, allow for considerable technique manipulation (such as lowering kVp, reducing scatter, subtracting one film from another) while staying within the exposure constraint of a few R for computed tomography. It becomes clear, however, that the utility of such techniques will depend not only on the results of the basic signal to noise levels; but also, on the extent of influence on the pattern recognition and structure noise aspects of the actual resultant clinical image.

APPENDIX I

Contrast In Radiography

$$D = G \log_{10} E + K = -\log_{10} T$$
$$\Delta D = 0.43G \frac{\Delta E}{E} = \frac{-0.43\Delta T}{T}$$

Then, for $C_R = \frac{T_S - T_B}{T_B} = \frac{-G\,(E_S - E_B)}{E_B} = \frac{D_S - D_B}{0.43}$

When following the terminology used in the text, where E_S is the exposure incident on the screen-film combination in the signal region;

$$E_S = E_o\, e^{-\mu_B(x-d)}\, e^{-\mu_S d} = E_o\, e^{-\mu_B x}\, e^{-(\mu_S \mu_B)d}$$

and $E_B = E_o\, e^{-\mu_B x}$

Therefore, $C_R = -G\left(\frac{E_S - E_B}{E_B}\right) = G\left(1 - e^{-(\mu_S - \mu_B)d}\right)$

$$C_R \doteq G\,(\mu_S - \mu_B)\,d \qquad \text{without scatter}$$

$$C_R \doteq G\,(\mu_S - \mu_B)\,d\,\frac{E_B}{E_B + E_\phi} \qquad \text{with scatter}$$

APPENDIX II

Contrast In Computed Tomography

$$\xi = CT\# = K_1\,\frac{(\mu - \mu_w)}{\mu_w}$$

$$T = \text{luminance function} = K_2\,(\xi - \xi_L) = K_3\,(\mu - \mu_L)$$

ξ_L = CT# of lower window setting (μ_L)

Then $$C_{CT} = \frac{T_S - T_B}{T_B} = \frac{\mu_S - \mu_B}{\mu_B - \mu_L}$$

Where μ_B = linear attenuation coefficient of background

μ_S = linear attenuation coefficient of signal

μ_L = linear attenuation coefficient corresponding to lower window setting

μ_w = linear attenuation coefficient to water

K_1, K_2, K_3 = Constants

$$C_{CT} = \frac{\mu_S - \mu_B}{\mu_B - \mu_L}$$

REFERENCES

1. H. R. Blackwell, "Contrast thresholds of the human eye," J. Opt. Soc. Amer. 36, 624, 1946.

2. D. J. Goodenough, "Assessment of Image Quality of Diagnostic Imaging Systems," Medical Images: Formation, Perception and Measurement,, Proceedings of the 7th L. H. Gray Conference, Leeds, England, April 1976.

3. H. E. Johns and J. R. Cunningham, The Physics of Radiology, Charles C. Thomas, Springfield, Ill., 1953.

4. Michel M. Ter-Pogossian, The Physical Aspects of Diagnostic Radiology, Harper and Row, New York N.Y. 1967.

5. G. T. Barnes, H. Cleare and I. A. Brezovich, "Reduction of Scatter in Diagnostic Radiology by means of a scanning multiple slit assembly," Radiology, 120: 691, 1976.

6. E. C. McCullough, H. I. Baker, O. W. Hauser and D. F. Reese, "An Evaluation of the Quantitative and Radiation Features of a Scanning X-Ray Transverse Axial Tomograph: The EMI Scanner," Radiology, 111:709-715, June 1974.

7. T. R. Fewell and K. E. Weaver, "The Measurement of Diagnostic X-Ray Spectra with a High Purity Germanium Spectrometer," Proceedings of the 1974 BRH/SPIE Symposium - Medical X-Ray Photo-Optical Systems Evaluation, Published by SPIE, February 1976, Vol. 56.

8. M. E. Phelps, M. H. Galdo and E. J. Hoffman, "Correlation of Effective Atomic Number and Electron Density with Attenuation Coefficients Measured with Polychromatic X-Rays," Radiology, 117:585-588, December 1975.

9. M. M. Ter-Pogossian, M. E. Phelps, G. L. Brownell, J. R. Cox, Jr., D. O. Davis, and R. G. Evans, Reconstruction Tomography in Diagnostic Radiology and Nuclear Medicine, University Park Press, Baltimore, Md., 1977.

10. D. J. Goodenough, K. Rossmann and L. B. Lusted, "Factors Affecting the Detectability of a Simulated Radiographic Signal," Radiology, 8, 5:339-344, Sept.-October 1973.

11. K. Rossmann, "Spatial Fluctuations of X-Ray Quanta and the Recording of Radiographic Mottle," American Jour. of Roentgenology, 90:863-869, October 1963.

12. D. J. Goodenough, "Objective Measures Related to ROC Curves," Proceedings of the Society of Photo-Optical Instrumentation Engineers Symposium, Kansas City, 1974.

13. E. Tanaka and T. A. Iinuma, "Correction Functions for Optimizing the Reconstructed Image in Transverse Section Scan," Phys. Med. Biol., 20, 5:789-798, 1975.

14. K. Rossmann, "Point Spread Function, Line Spread Function and Modulation Transfer Function: Tools for the Study of Imaging Systems," Radiology, 93:257-272, August 1969.

15. K. E. Weaver, R. F. Wagner and D. J. Goodenough, "Performance Considerations of X-Ray Tube Focal Spots," Proceedings of the 1974 BRH/SPIE Symposium - Medical X-Ray Photo-Optical Systems Evaluation. Published by SPIE, Feb. 1976, 56.

16. D. J. Goodenough, K. E. Weaver and D. O. Davis, "Development of a Phantom for Evaluation and Assurance of Image Quality in CT Scanning," Optical Engineering, 16, 1:52-57, January-February 1976.

DISCUSSION:

Gopala Rao expressed agreement with the presented results and noted that the size of the object to be imaged has a great influence upon the type of recording medium to be used. The participants agreed that conventional x-ray film is best for imaging relatively high-contrast objects, even though some image degradation does occur when three-dimensional objects are recorded on two-dimensional film. CT images are best used to image objects with relatively low-contrast properties and only moderate spatial resolution constraints. Imaging of objects requiring high spatial resolution (e.g., microcalcification) is best done with film with high spatial resolution properties and lower contrast resolution properties (e.g., xeroradiographs).

Brian Pullan began a general discussion regarding the relationship between the signal-to-noise ratio and the object size resolution. All participants agreed that the relationship is complex, being both a function of ambient background noise and the perception capabilities of the human-eye-brain combination. Dr. Goodenough felt that on an equivalent radiation dose basis, x-ray film and CT images both have comparable and well-defined relationships which relate image contrast and spatial resolution to the inherent image noise levels. He feels that anomalous structures in CT images can be perceived better than with conventional x-ray techniques. He also noted that nuclear medicine images have a higher S/N ratio, despite a lower spatial resolution capacity.

THREE-DIMENSIONAL RECONSTRUCTION IN LONGITUDINAL TOMOGRAPHY BY MEANS OF ITERATIVE APPROXIMATION

T. Tomitani and E. Tanaka
National Institute of Radiological Sciences
Anagawa, Chiba-Shi, Japan

ABSTRACT

Longitudinal tomographic devices suffer from the disadvantage that the off-focus images are superimposed on the in-focus image. In an effort to eliminate these off-focus images from the observed sectional images, here is presented an iterative approximation method similar to the iterative approximation method known as Gauss-Seidel method in linear equation theory. The condition of convergence of the present iterative approximation method was analyzed and a method to prevent apparent divergence of the iteration process in the low frequency region is presented. The present iterative method was tested on a simulation phantom of liver and pancreas, which showed that 2-4 iterations seem to be sufficient for practical purposes.

INTRODUCTION

Recently emission type transverse axial tomography has attracted special interest. These types of tomographic devices, especially those which utilize positrons, give us good quantitative information, although slice width is limited to 1~2.5 cm at present. Besides this, there are some devices which give us depth information. The following devices can be used for longitudinal tomography, i.e., the tomographic scanning camera type devices, the scintillation camera with rotating inclined hole collimator, coded aperture imaging and the positron camera which utilizes two scintillation cameras or proportional counters. These devices produce poor quantitative information as to depth direction, in the sense that the absolute intensities on the sectional planes perpendicular to the detector axis cannot be determined. However these devices produce images from multiple regions on the long axis of the body rather than a single transverse slice and are suitable for imaging large organs. Therefore, these two transverse types of tomography complement each other in practice.

Longitudinal tomographic devices suffer from such disadvantages that quantitative information cannot be obtained and that the off-focus images are superimposed on the in-focus images. The resultant images may be even worse than the images obtained by conventional scintillation cameras.

Several investigations have been carried out to eliminate these off-focus

images. Schmidlin studied the iterative search algorithm,[1] which is similar to Algebraic Reconstruction Technique in transverse axial tomography. Myers studied the method, in which adjacent images are blurred by an amount equal to the response between two planes and are subtracted from the images to eliminate the off-focus images on the adjacent planes.[2] Chang, Macdonald and Perez-Mendez studied Fourier methods.[3] In the frequency domain, the equation in question transformed into a set of linear equations which can be solved provided the determinant of the coefficient matrix is not zero. Chu and Tam treated this problem in such a way that they utilized weighted Green's function and deconvoluted the observed images by Fourier method.[4]

Here is presented an iterative method to eliminate the off-focus images from the observed images. Since this method can be explained more clearly by analogy with linear equation, a brief explanation on linear equation theory is given in Sec.2-1. In Sec. 2-2, the iterative method applied to this problem is presented and in Sec. 2-3 the condition of convergence of the present iterative method is presented. In Sec. 3, the results on some simulation studies are given.

ITERATIVE APPROXIMATION METHOD

Gauss-Seidel's Iterative Approximation Method in Linear Equation Theory

Linear equations are represented by eq. (1), where X,F and A denote unknown vector, constant vector and coefficient matrix, respectively.

$$AX = F \tag{1}$$

This equation can be solved by means of Gauss-Seidel iterative method, when the absolute values of diagonal components of A are considerably larger than the absolute values of off-diagonal components of A.[5] Iterative method follows the steps shown in eq. (2).

$$X^{(k)} = X^{(k-1)} + H^{(k)}(F - AX^{(k-1)}) \tag{2}$$

which generates vector series $X^{(1)}, X^{(2)}, \ldots, X^{(k)}, \ldots$, where $H^{(1)}, H^{(2)}, \ldots, H^{(k)}, \ldots$ are a fixed series of matrices, and $X^{(0)}$ is an initial approximate vector of X. This initial vector $X^{(0)}$ is, in general, arbitrary, but the choice of $X^{(0)}$ affects the speed of convergence of eq. (2). In practice, F itself is chosen as $X^{(0)}$. Thus k-th approximate vector can be obtained. True solution vector of eq. (1) is invariant with respect to the steps of eq. (2). A proper choice of a matrix series $H^{(k)}$ enhances convergence of iteration, or, in some cases, prevents divergence. If $H^{(k)} = H$ (k=1,2,....) this process is called a stationary process. In this case, eq.(1) can be transformed into equivalent eq. (3),

$$HAX = F' \tag{3}$$

or

$$A'X = F' \tag{3'}$$

where A' = HA and F' = HF. With this notation, eq. (2) is transformed to

$$X^{(k)} = X^{(k-1)} + F' - A'X^{(k-1)} \tag{4}$$

or

$$X^{(k)} = (E - A')X^{(k-1)} + F' = BX^{(k-1)} + F' \tag{4'}$$

where B = E - A'.

Then $X^{(k)}$ can be represented by eq. (5).

$$X^{(k)} = B^k X^{(0)} + \sum_{j=0}^{k-1} B^j F' \tag{5}$$

The necessary and sufficient condition of convergence of a series in the right hand of eq. (5) is "all absolute eigenvalues of matrix B should be less than 1."[5] The simple iterative method is to choose H = E, then B = E - A. The above mentioned condition is equivalent to "all eigenvalues of A should be positive and less than 2." If the eigenvalues exceed 2, the iteration will diverge. By the choice of $H = 2/\mu \cdot E$, where μ is the upper limit of eigen value of A, we can make absolute eigenvalues of $E - HA = E - 2/\mu A$ less than 1, so that a series in the right hand of eq. (5) will converge. This factor μ controlls the degree of convergence, so that it can be regarded as "damping factor." If the eigenvalues are small, smaller value of μ can be adopted, which enhances the speed of convergence. If the eigenvalues are large, we have to use large μ so as to converge the series in eq. , but at a cost of convergence speed.

Application of Iterative Approximation Method to Longitudinal Tomography.

Here let us consider 5 sectional planes as shown in Fig. 1. A_i (i=1~4) denotes the blurring operator, where index denotes the separation between two planes. I_i denotes the "true" image on the i-th plane and F_i denotes the observed image focused to i-th plane. Then the following equation holds:

$$\sum_{j=1}^{5} A_{|i-j|} * I_j = F_i \quad (i = 1,2,\ldots,5) \tag{6}$$

in which the operation of the type $A_i * I_j$ means two dimensional convolution. Note that the blurring operator depends only on the separation between two planes, therefore $|i-j|$ characterizes the blurring operator. Eq. (6) is formally analogous to eq. (1). The essential features of iteration method lie in the following steps.[6,7]

(a) assume an initial approximate solution,
(b) calculate error terms as shown in the second term of the right hand in eq. (2), no matter what arithmetic operation it may be,
(c) correct the approximate solution to obtain better approximation of the solution,
(d) go to step (b).

Therefore, this method can be applicable to the problem of eq. (6), provided that the operations of blurring A_i's fairly spread out, which ensures the convergence of iteration. The calculation algorithm is essentially identical to eq.(2), except that elements of A should be replaced by blurring operator $A_i(\underset{\sim}{r})$, vector X by a set of matrices and simple products should be replaced by convolution. As to first step (a), we usually adopt F_i's themselves. If we choose $H = 2/\mu \cdot E$ so as to prevent divergence, then the k-th iteration process is:

$$X_i^{(k)} = X_i^{(k-1)} + \frac{2}{\mu}\{F_i - \sum_{j=1}^{5} A_{|i-j|} * X_j^{(k-1)}\} \tag{7}$$

The damping factor, μ, will be determined later. Note that the total counts of the approximate solution $X_i^{(k)}$ is invariant through the iteration process, and is equal to the total counts of $X_i^{(0)}$, because the total counts of the terms in the brackets are 0. The longitudinal section imaging gives us no information of the D.C. components of the section and little information on low frequency components, because this method can be regarded as incomplete projection, in which the angle of measurements with respect to the axis of the detector(s) is limited to 0°~40° in practical cases.[3] Usually F_i's are adopted as $X_i^{(0)}$'s, so that the total counts of F_i's are preserved throughout the process. If some elements of $A_{|i-j|} * X_j^{(k-1)}$ overflow out of the field of view, a proper weight should be multiplied to $A_{|i-j|}$ so as to normalize the summation of $A_{|i-j|}$ to unity.

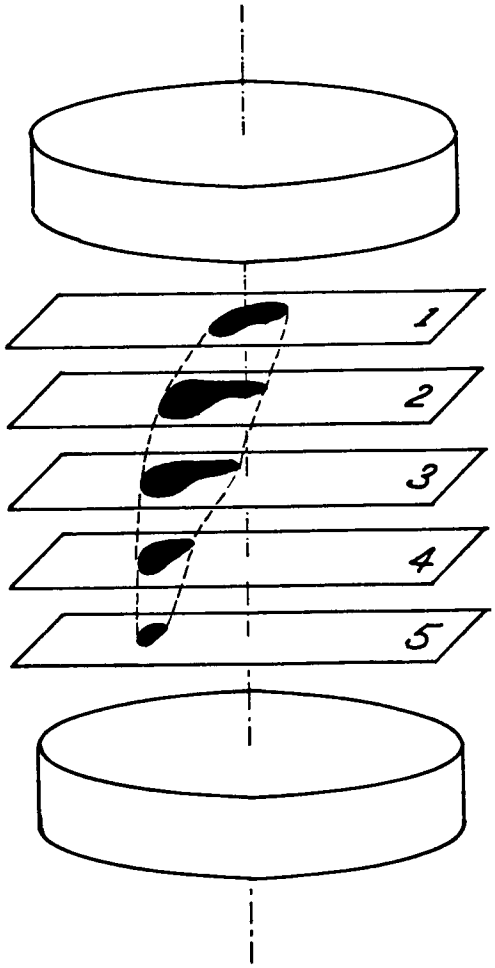


Fig. 1. Geometrical Layout of Longitudinal Tomography.

Condition of Convergence of Iterative Approximation Method

The condition of convergence can be found more clearly in frequency space. Fourier transformation of eq. (6) leads to eq. (8).

$$\sum_{j=1}^{5} A_{|i-j|}(\omega)\cdot I_j(\omega) = F_i(\omega) \quad (i=1,2,\ldots\ldots,5) \tag{8}$$

Convolutions between $A_{|i-j|}$'s and I_j's in eq. (6) are replaced by simple products of scalar numbers and eq. (8) is a simple set of linear equations.[4] Then the condition of convergence of the iterative approximation is "all eigenvalues of A should be positive and less than 2" as stated earlier. The condition of convergence is therefore frequency-dependent.

Here we assume the form of A_i's as eq. (9) (Fig. 2 a and b), in which A's are axially symmetric and rectangular in radial direction.

$$A_k(r) = \begin{cases} \dfrac{1}{4\pi R_k t_k} & (R_k - t_k \leq r \leq R_k + t_k) \\ 0 & (\text{otherwise}) \end{cases} \tag{9}$$

In eq. (9), R_k and $2t_k$ denote the mean radius and the thickness of the "ring," respectively. This type of the blurring operators simulates the condition encountered in positron cameras.[4] In the case of positron cameras it is advantageous to eliminate gamma-rays of small angle with respect to the detector axis, because such information has less depth information.

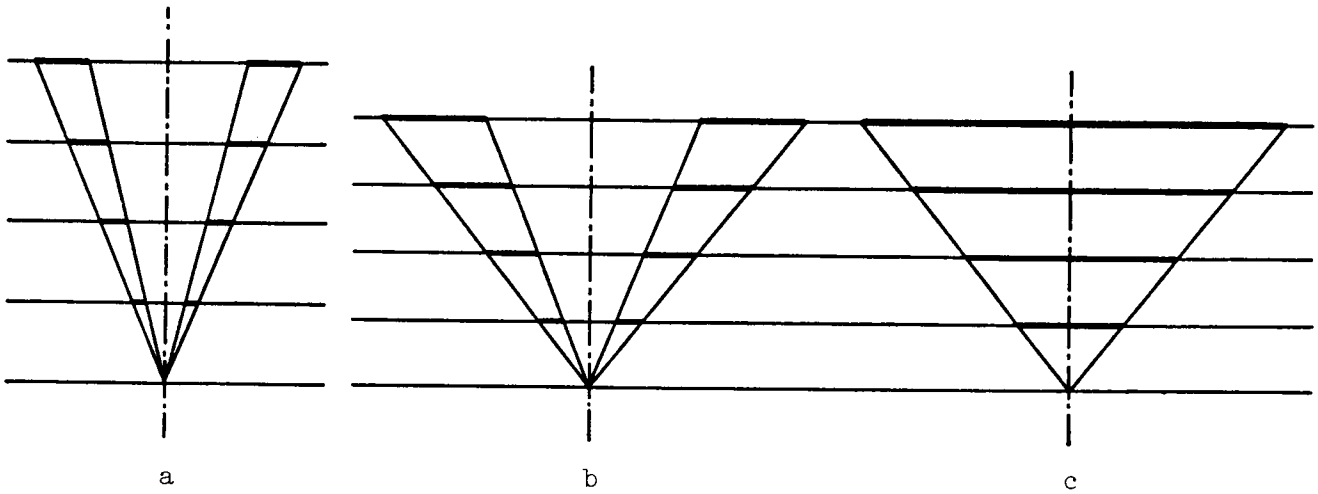


Fig.2. Geometrical Conditions of Various Blurring Operators.

In the case of the tomographic camera with rotating inclined hole collimator, t_k's are considerably smaller than R_k's, yet eq. (9) can be used.

Fourier transformations of eq. (9) are;

$$A_k(\omega) = \int_0^\infty A_k(r)J_0(r\omega)rdr = \int_{R_k-t_k}^{R_k+t_k} \frac{rJ_0(r\omega)}{4\pi R_k t_k} dr$$

$$= \frac{1}{4\pi R_k t_k \omega} \{rJ_1(r\omega)\}_{R_k-t_k}^{R_k+t_k} \qquad (10)$$

where $J_i(x)$ denotes i-th Bessel's function. When sectional planes are equally spaced, spacing, s_k, is equal to ks, where s is the separation between two adjacent planes, and $R_k=kR_1$, $t_k=kt_1$. Eigenvalues of matrix $A(\omega)$ as a function of angular frequency can be calculated with these $A_k(\omega)$ (k=1,2,3,4). Two numerical examples are shown in Fig. 3 a and b as to the operators shown in Fig. 2 a and b. As to a in Fig. 2, s=3cm, R_1=1cm and t_1=0.25cm are assumed. As to b in Fig. 2, s=2.5cm, R_1=1.5cm and t_1=0.5cm are assumed. In Fig. 3 b, all eigenvalues of angular frequency higher than ~3.5 are almost 1 which means that matrix $A(\omega)$'s are almost equal to identity matrix, E, in such frequency region. In Fig. 3, eigenvalues exceed 2 at lower frequency region, if we choose H = E. Since maximum eigenvalue of $A(\omega)$ is 5, it seems reasonable to select $\mu = 5$, i.e. $H = 2/5 \cdot E$. Also it is apparent from Fig. 3 a and b that the latter case is more stable at lower frequency than the former. As to the example a, cone angle is between 14°~23° and as to the example b, cone angle is between 22°~39°.

Tomographic effect is better with larger cone angle, but at the expense of the spatial resolution. Cone angle should be determined by trading-off between spatial resolution and tomographic effect. Here the latter geometric condition was adopted in the simulation study.

The following scalar associated with matrix A is considered as a measure of instability of a linear equation.[5] Here instability means that small changes either in the coefficients or in constant terms in a linear equation lead to

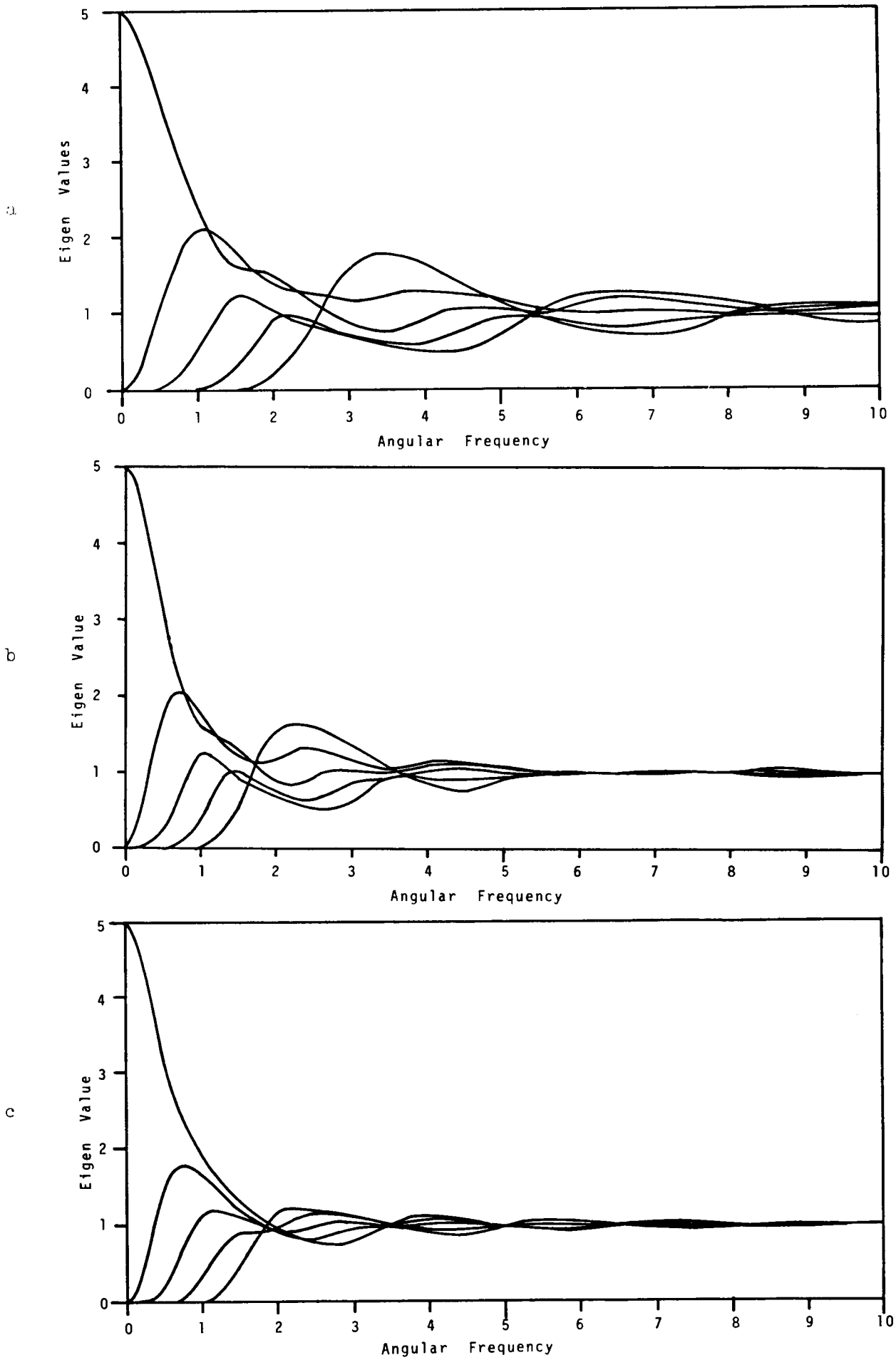


Fig. 3. Eigenvalues of the Operators shown in Fig. 2.

larger variations in the solution.

$$\nu(A) = \frac{1}{n}\sqrt{\left(\sum_{i=1}^{n} \mu_i\right)\left(\sum_{i=1}^{n} 1/\mu_i\right)} \tag{11}$$

Here μ_i's$(i=1,2,\ldots,n)$ denote the **eigenvalues** of a matrix A. $\nu(A)$'s are shown in Fig. 4 as a function of ω in case of the blurring operators used in the examples a and b in Fig. 3. This number is equal to or larger than 1. The larger $\nu(A)$ is, the more unstable the solution is.

For comparison, eigen**values** of "disc" type operators shown in Fig. 2c were calculated and are shown in Fig. 3c. The radii of the discs were set equal to the outer radii of the ring type operators shown in Fig. 2b. Although eigen**values** of disc type operators are a little smaller than those of ring type operators, but $\nu(A)$ of the former is a little larger than $\nu(A)$ of the latter below angular frequency of ~ 2. Anyway the difference between them is not significant. Hence disc type operators are essentially as stable as ring type operators.

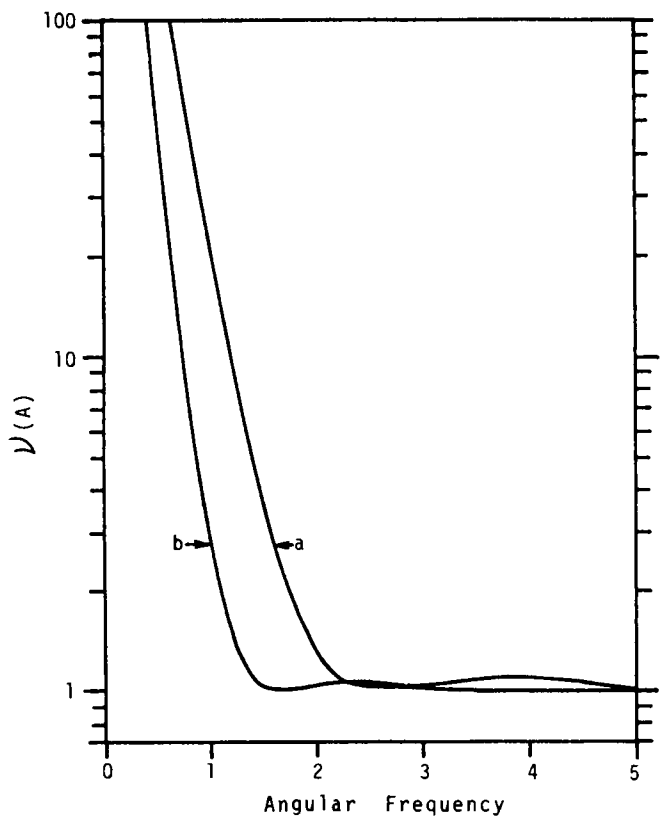


Fig. 4. $\nu(A)$'s of the Operators of a and b in Fig. 2.

The Recurring Iteration Method versus the Successive Iteration Method.

The approximate solution of i-th plane in k-th iteration can be written explicitly as;

$$\begin{aligned} X_i^{(k)} &= X_i^{(k-1)} + H\left(F_i - \sum_{j=1}^{5} A_{|i-j|} * X_j^{(k-1)}\right) \\ &= X_i^{(k-1)} + H\left(F_i - \sum_{j=1}^{i-1} A_{|i-j|} * \underline{X_j^{(k-1)}} - \sum_{j=i}^{5} A_{|i-j|} * X_j^{(k-1)}\right) \end{aligned} \tag{12}$$

Up to this step, $X_j^{(k)}$'s ($1 \leq j < i$) have been already calculated. Therefore it is possible to replace $X_j^{(k-1)}$'s by $X_j^{(k)}$'s for $1 \leq j < i$ (underlined term in eq. (12)). We can save memory with this algorithm, because $X_i^{(k)}$ can be stored in the same memory location of $X_i^{(k-1)}$. This **algorithm is known as "recurring" iteration** method in the linear equation theory, while the method in Sec.2-1 is known as "successive" iteration method.

In order to examine goodness of approximation, we adopted "distance" Θ defined by eq. (13).

$$\Theta = \arccos\left\{\frac{(X,I)}{|X| \cdot |I|}\right\}/(\pi/2) \tag{13}$$

Images in N^2 dimensional space can be regarded as vectors. In eq. (13), X and I denote the approximate solution and the true image, respectively and (X,I) denotes scalar product of them. $|X|$ and $|I|$ denote $\sqrt{(X,X)}$ and $\sqrt{(I,I)}$. Therefore distance Θ thus defined can be interpreted as an angle between X and I. For simplicity, angle is measured in unit of radian/($\pi/2$). If two images exactly coincide each other, the angle is 0.

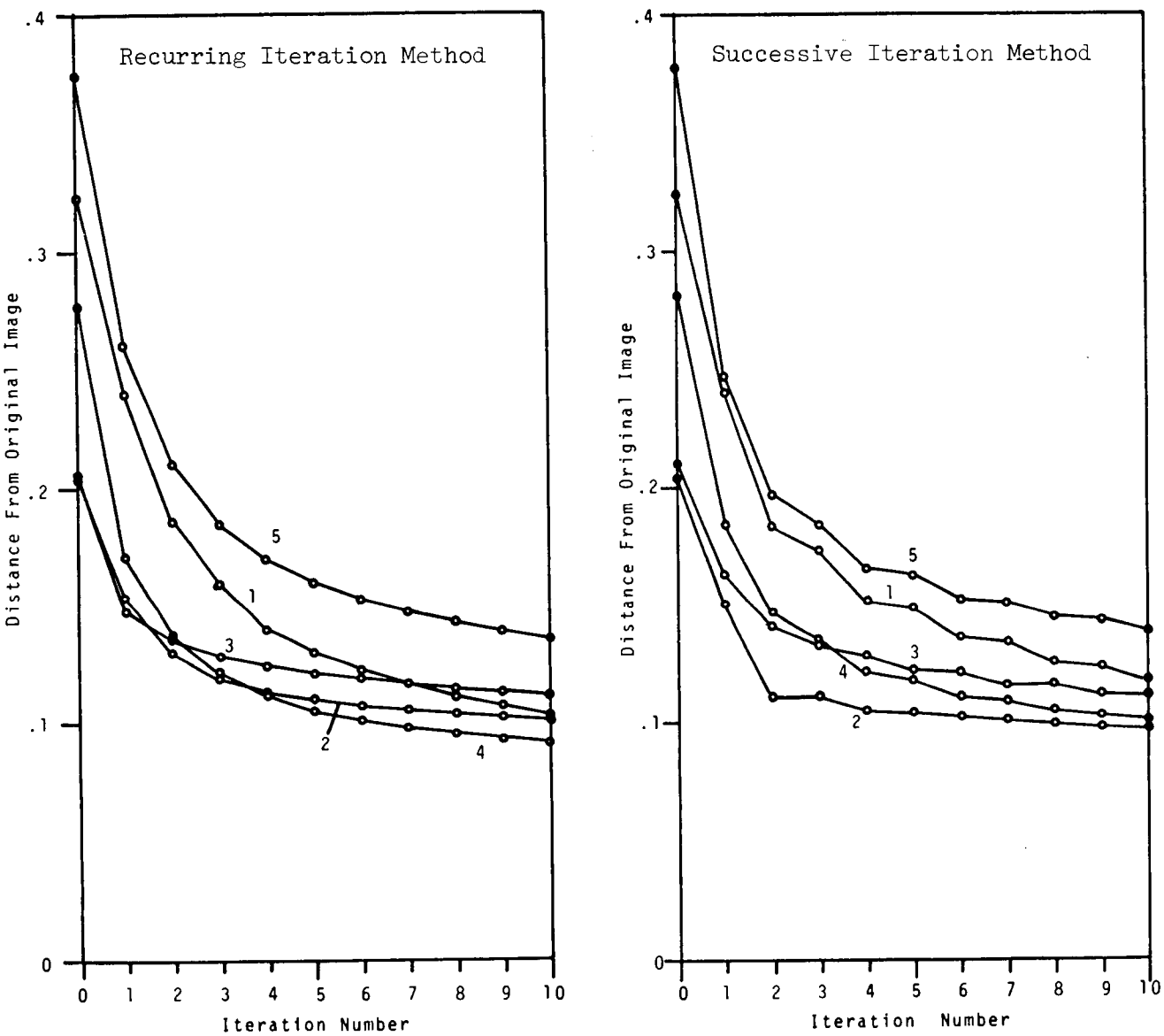


Fig. 5. Comparison of Convergence Characteristics between Successive Iteration Method and Recurring Iteration Method.

The convergence characteristics of the recurring iteration method is compared with that of the successive iteration method in Fig. 5, with the simulated phantom similar to Fig.9. In this case $\mu=5$ was adopted as a damping factor. From Fig. 5, the successive iteration method shows oscillation. Yet, the characteristics of convergence of this method is essentially equal to that of the recurring iteration method. Because of simplicity in algorithm, the recurring iteration method was adopted in this simulation study.

Convolution of Operators

Since iteration processes involve convolutions of the type $A_i * A_j$, it is worth while to analyze the general behavior of this type of convolution.

$$A_i * A_j(r) = 2\pi\int_0^\infty A_i(\omega)\cdot A_j(\omega)J_0(r\omega)\omega d\omega$$

$$= 2\pi\int_0^\infty \frac{1}{4\pi R_i t_i \omega}\{rJ_1(r\omega)\}_{R_i-t_i}^{R_i+t_i} \frac{1}{4\pi R_j t_j \omega}\{rJ_1(r\omega)\}_{R_j-t_j}^{R_j+t_j} J_0(r\omega)\omega d\omega \qquad (14)$$

which involves the following integral,[8]

$$I = \int_0^\infty \frac{J_1(a\omega)J_1(b\omega)J_0(r\omega)}{\omega} d\omega = -\frac{2}{\pi ab}\int^r \frac{\Delta(a,b,r)}{r} dr + \text{constant} \qquad (15)$$

where $\Delta(a,b,r)$ is the area of the triangle surrounded by sides a,b and r.

$$I = \frac{1}{4\pi ab}\{ -\sqrt{-r^4+2(a^2+b^2)r^2-(a^2-b^2)} + (a^2+b^2)\sin^{-1}\left(\frac{-r^2+a^2+b^2}{ab}\right)$$

$$+ |a^2-b^2|\sin^{-1}\left[\frac{(a^2+b^2)r^2-(a^2-b^2)^2}{2abr^2}\right]\} + \frac{b}{4a} \qquad (15')$$

Several examples of these convolutions are shown in Fig. 6. Myers[2] assumed that "A_i is linearly related to the plane idex i such that $A_i = A_1^i$ and $A_i * A_j = A_{i+j}$." From Fig. 6, it is apparent that his assumption does not hold even approximately in this case. For example, A_2 is not equal to $A_1 * A_1$. Nevertheless his method to eliminate off-focus blurred images is a good first order approximation. Let us examine the third plane, for example.

$$\begin{cases} A_1*I_1 + I_2 + A_1*I_3 + A_2*I_4 + A_3*I_5 = F_2 \\ A_2*I_1 + A_1*I_2 + I_3 + A_1*I_4 + A_2*I_5 = F_3 \\ A_3*I_1 + A_2*I_2 + A_1*I_3 + I_4 + A_1*I_5 = F_4 \end{cases} \qquad (16)$$

In F_3, I_3 is an in-focus image, while the rests are out-of-focus images of the other planes of which A_1*I_2 and A_1*I_4 are predominant. First let us apply A_1 to F_2 and F_4 and then subtract them from F_3, then the following equation results.

$$F_3 - A_1*F_2 - A_1*F_4 = (A_2 - A_1*A_1 - A_1*A_3)*I_1 - A_1*A_2*I_2 + (E - 2A_1*A_1)*I_3$$

$$- A_1*A_2*I_4 + (A_2 - A_1*A_3 - A_1*A_1)*I_5 \qquad (17)$$

In this equation, the terms A_1*I_2 and A_1*I_4, the blurred images on the adjacent planes are suppressed with the residues $A_1*A_2*I_2$, $A_1*A_2*I_4$ etc. which are much blurred compared to the images A_1*I_2 or A_1*I_4. This explains why Myers' method works well.

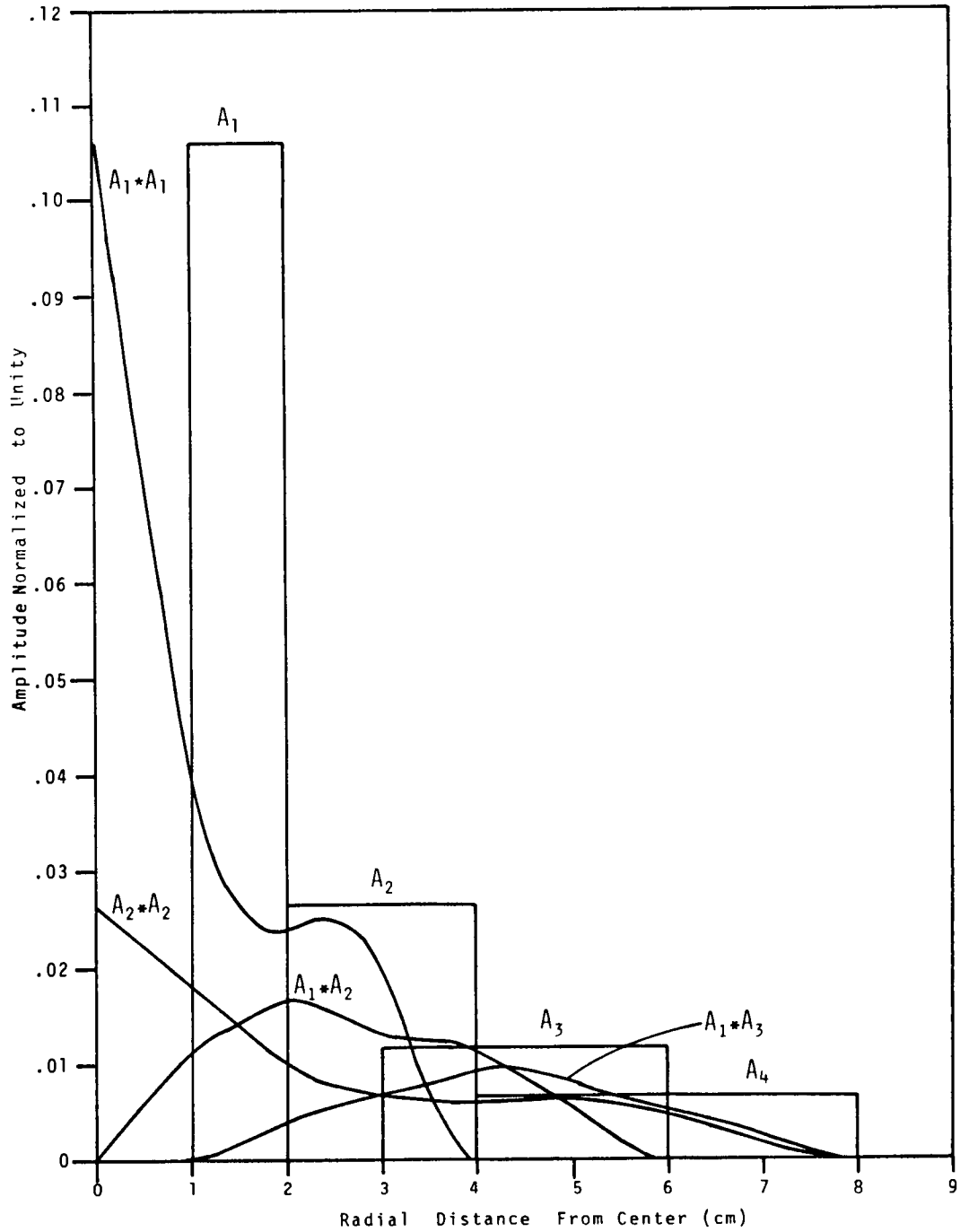
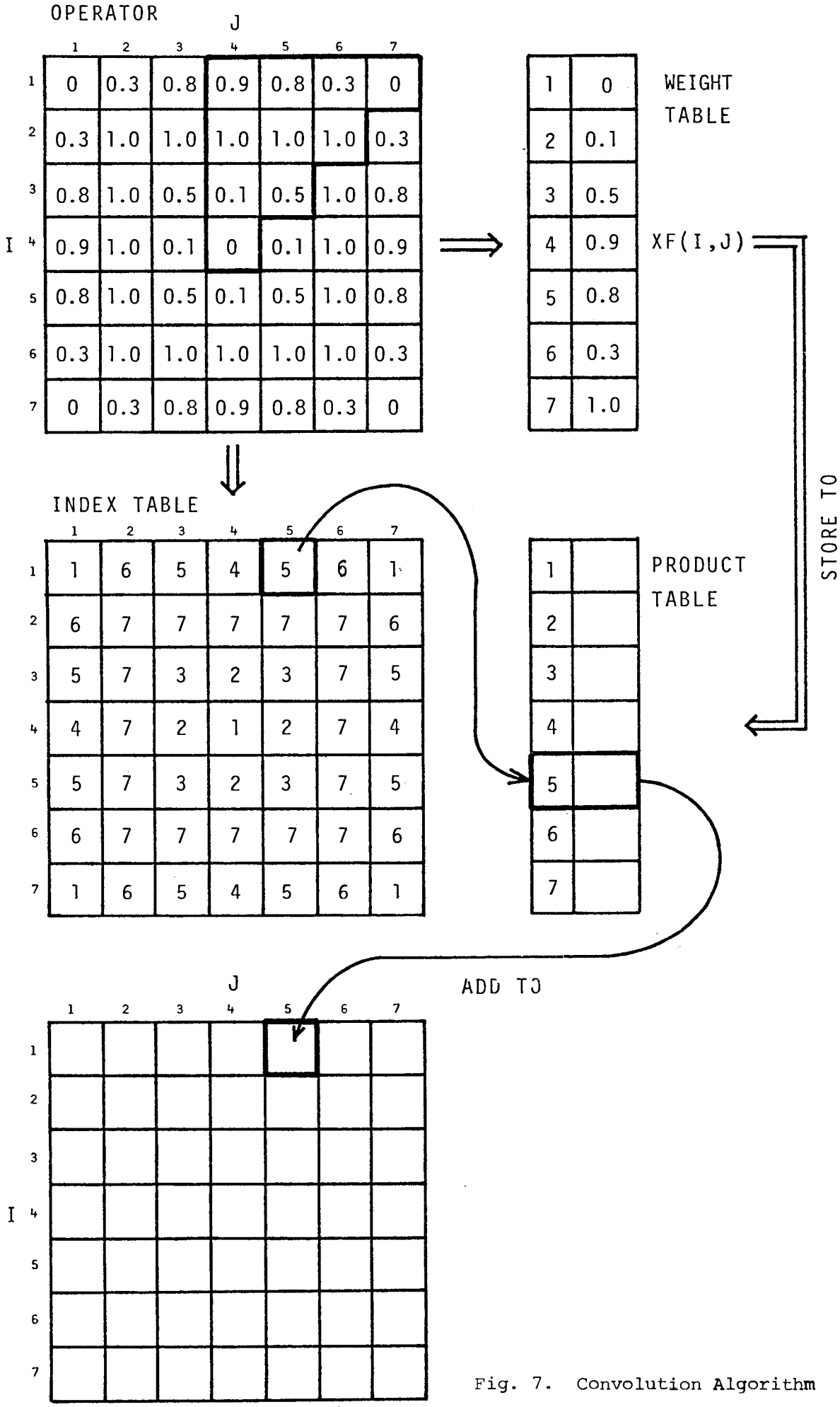


Fig. 6. Convolution of Operators.

A Convolution Algorithm to Reduce Calculation Time

In calculation of convolutions of the type A_i*F_j, simplification of calculation is achieved by taking account of the symmetries of operator A_i. In image processing, we often want to convolve two matrices, one of which has axial symmetry. Therefore, it seems worth while to consider this algorithm. In continuous space, A_i is symmetric with respect to the axis, but digitization process degrades symmetry, yet digitized A_i still has 4 axes of symmetry. By appropriate rearrangement of convolution procedures, the basic calculation procedure in FORTRAN can be expressed as;

$$H(I+M,J+N) = H(I+M,J+N) + A(M,N)*F(I,J) \qquad (18)$$



	1	2	3	4	5	6	7
1	0	0.3	0.8	0.9	0.8	0.3	0
2	0.3	1.0	1.0	1.0	1.0	1.0	0.3
3	0.8	1.0	0.5	0.1	0.5	1.0	0.8
4	0.9	1.0	0.1	0	0.1	1.0	0.9
5	0.8	1.0	0.5	0.1	0.5	1.0	0.8
6	0.3	1.0	1.0	1.0	1.0	1.0	0.3
7	0	0.3	0.8	0.9	0.8	0.3	0

1	0
2	0.1
3	0.5
4	0.9
5	0.8
6	0.3
7	1.0

	1	2	3	4	5	6	7
1	1	6	5	4	5	6	1
2	6	7	7	7	7	7	6
3	5	7	3	2	3	7	5
4	4	7	2	1	2	7	4
5	5	7	3	2	3	7	5
6	6	7	7	7	7	7	6
7	1	6	5	4	5	6	1

1	
2	
3	
4	
5	
6	
7	

Fig. 7. Convolution Algorithm

In eq. (18), we assume A(M,N) is symmetric. F(I,J) denotes one picture element to be convoluted and H(I+M,J+N) denotes the picture element to which the convoluted result is accumulated.

Since A(M,N) is symmetric, we need not calculate all products of A(M,N)·F(I,J), over all (M,N) space, but rather calculations of products over the basic area of (M,N) space are sufficient. In this way arithmetic multiplications can be reduced.

The necessary procedures are schematically shown in Fig. 7. Here we assume that the matrix size of A(M,N) is smaller than that of image matrix. Matrix A(M,N) is divided into two tables, i.e. "weight table" and "index table." In the weight table, A(M,N) values of basic unit are stored. We precalculate the products of W(K)·F(I,J) and store the results in the "product table." The index table size is equal to the size of A(M,N) and contains the index of the element in the product table into which the results are stored.

First we increase M,N and, in each step, the content of index table indexes into the product table and the content is read out and is added to image matrix H(I+M,J+N). In this way, the necessary product calculations can be reduced to 1/3~1/8, depending on the size of matrix A.

If there are the same matrix elements in A, then the products calculations can be reduced further. In the example shown in Fig.7, the net product calculations are reduced from 10 to 7. In total, 7x7=49 product calculations are necessary in the straightforward convolution calculation, while this algorithm reduces the net product calculations to 7. A reduction factor of 7 is achieved.

SIMULATION TESTS AND RESULTS

The present iterative approximation method was tested with a simulated phantom of liver and pancreas. Five planes were adopted and each picture element was assumed to 1 cm x 1 cm. Separation between two adjacent planes, s, was so chosen that the liver was divided into 5 slices and was set to 2.5 cm. The blurring operators shown in Fig. 2 b were adopted in this test. One plane consists of 32 x 32 picture elements.

In Fig. 9 and Fig. 13 a are shown the phantom used in this simulation study. In the displays Fig. 8 through Fig.12, one picture element is displayed as two characters in the lateral direction, so as to obtain an almost square display.

In Fig. 8 the conventional scintigram is shown for comparison which is the superposition of 5 images in Fig. 9.

In Fig. 10 the observed tomographic images are shown. Note that none of 3 cold spots are discernible in the conventional scintigram and that cold spots are barely discernible in the tomographic images in Fig. 10. Pancreas can be distinguished from liver neither in the conventional scintigram nor in the tomographic images.

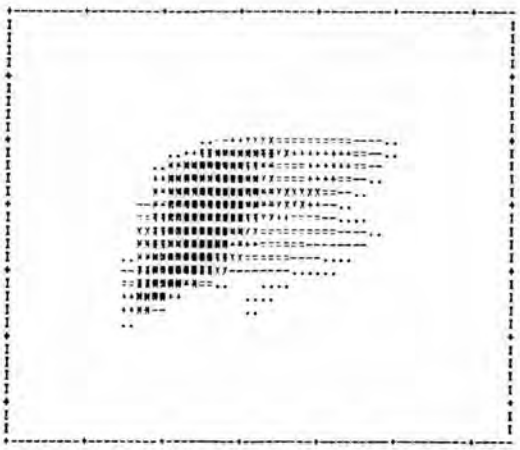


Fig. 8. Conventional Scintigram

In Fig. 11 are shown the restored sectional images without noise by means

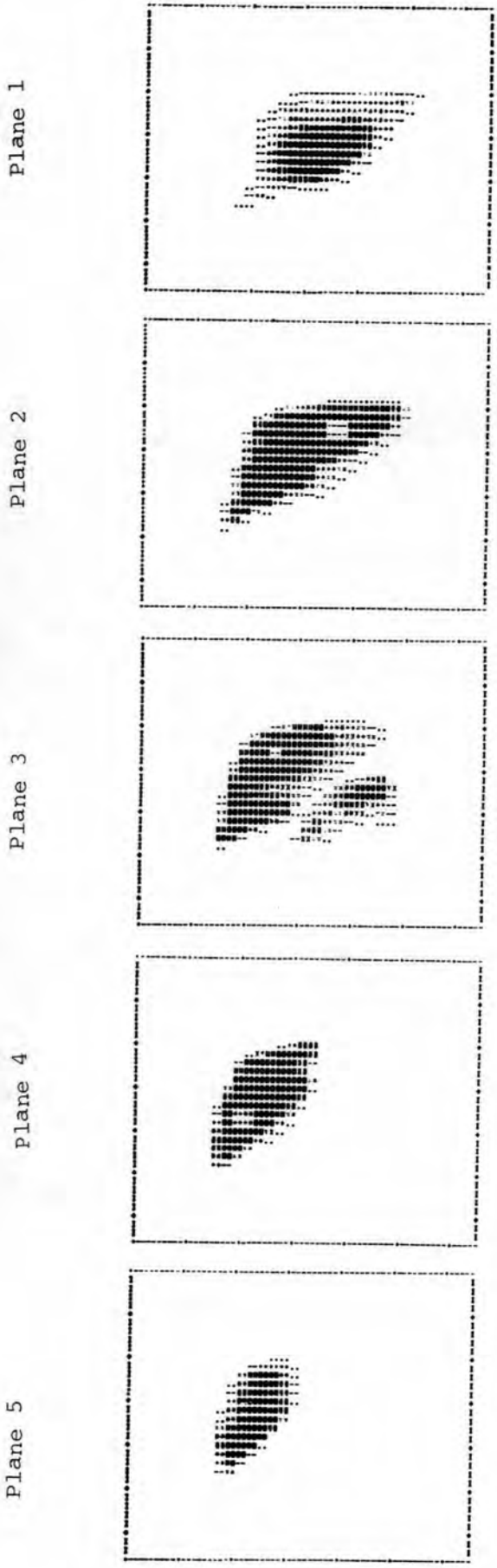


Fig. 9. Original Images.

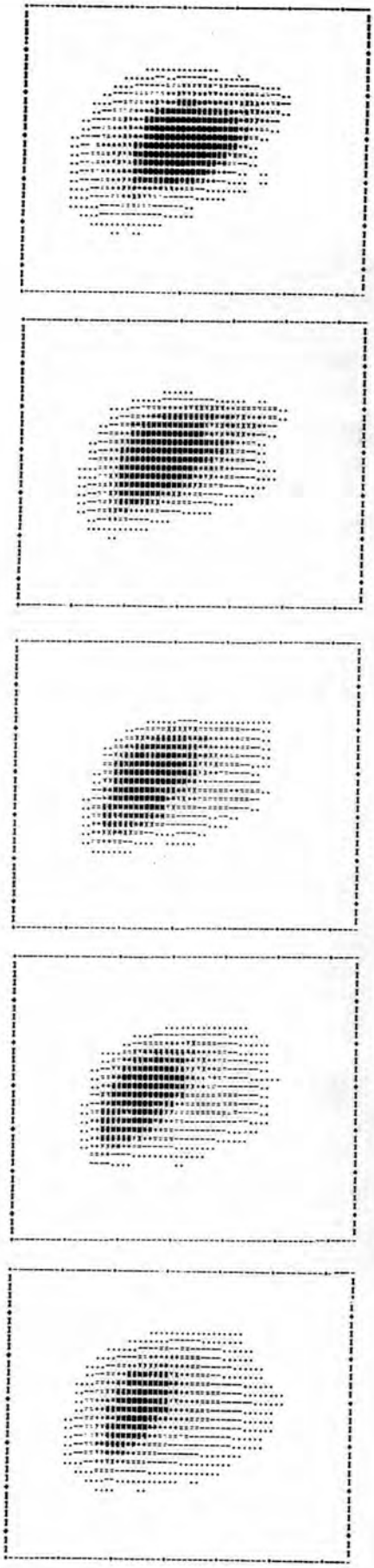
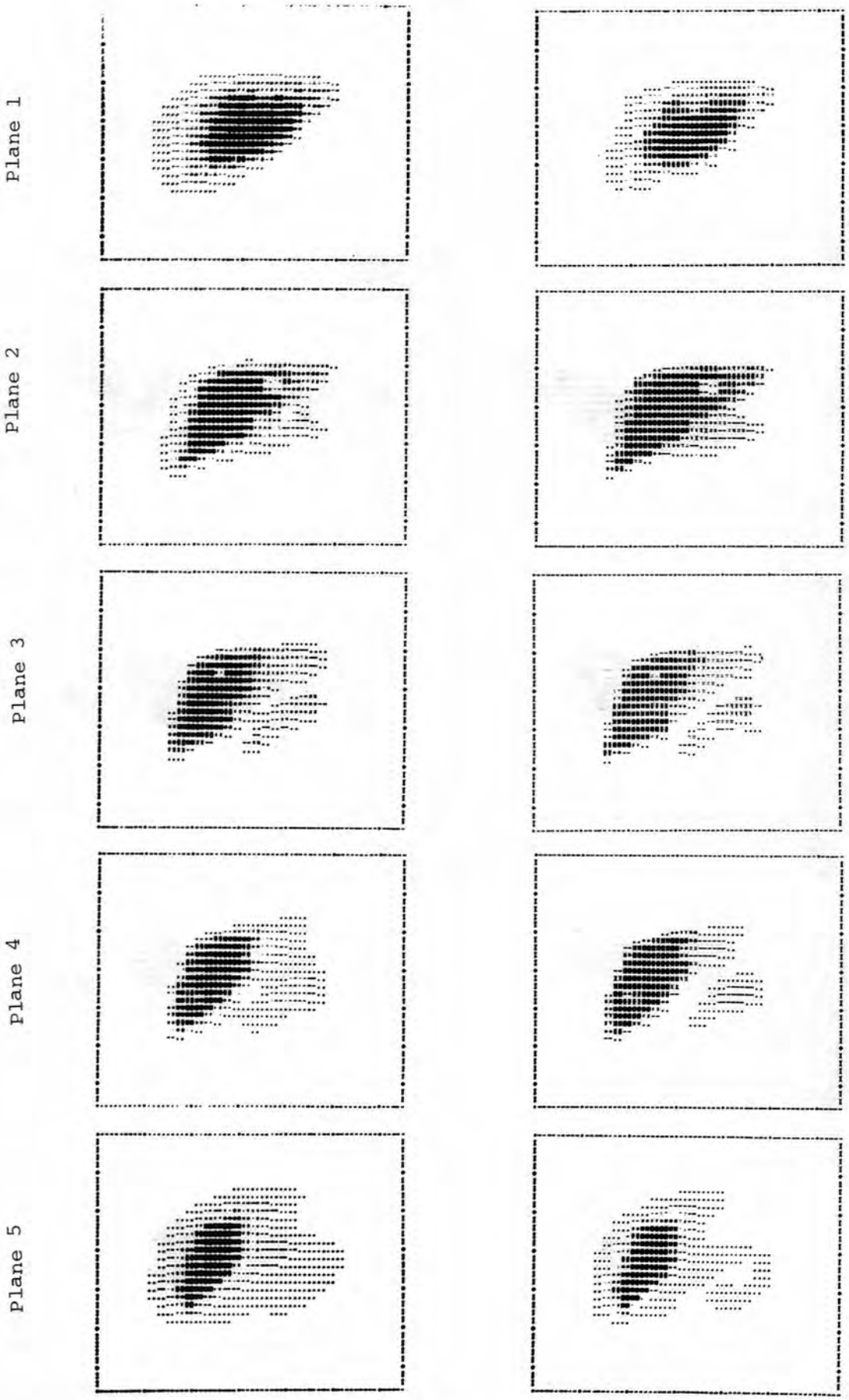
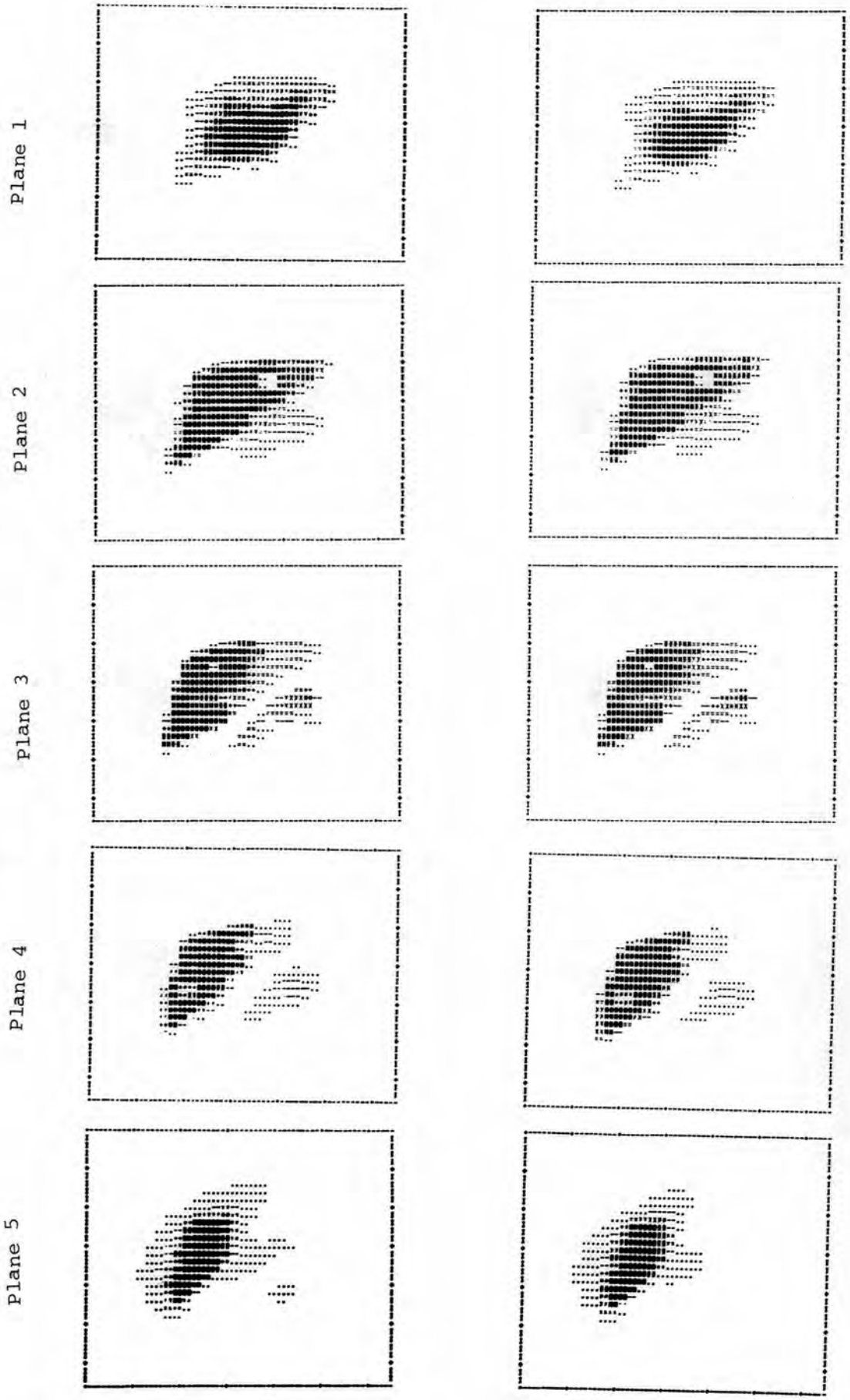


Fig. 10. Ohserved Images.



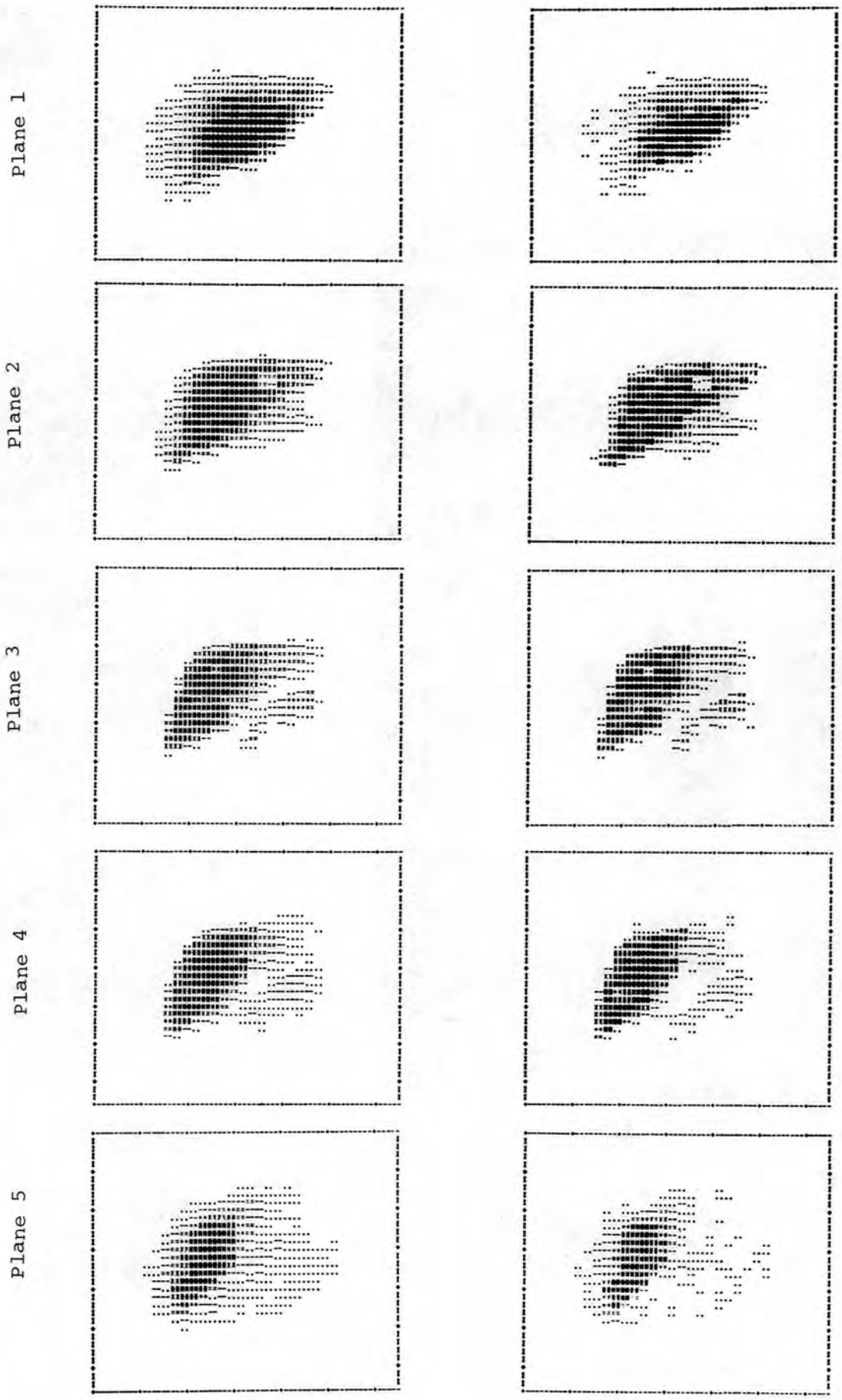
First Iteration Third Iteration

Fig. 11. Restored Images (without noise).



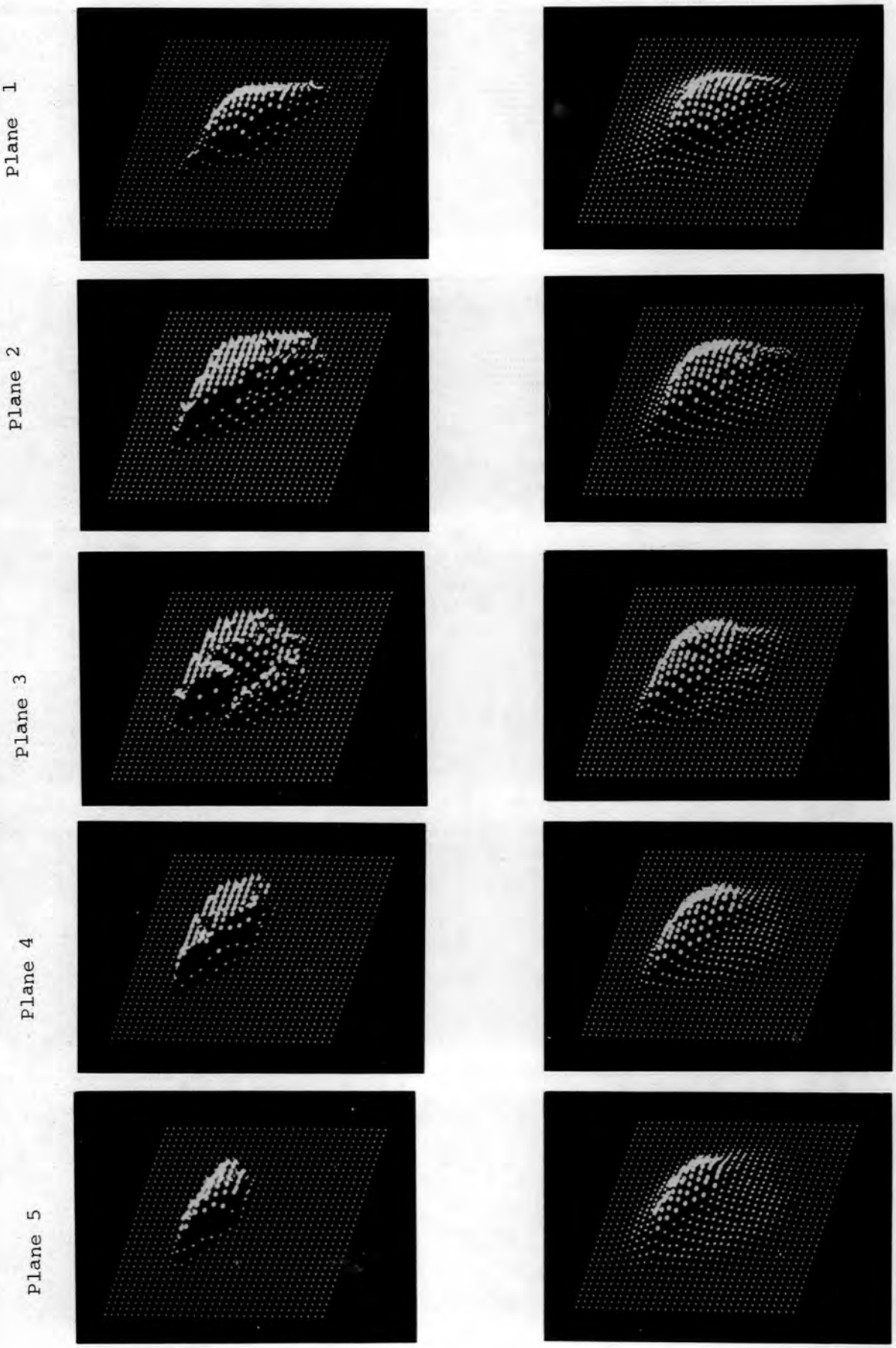
5-th Iteration 10-th Iteration

Fig. 11. Restored Images (without noise).



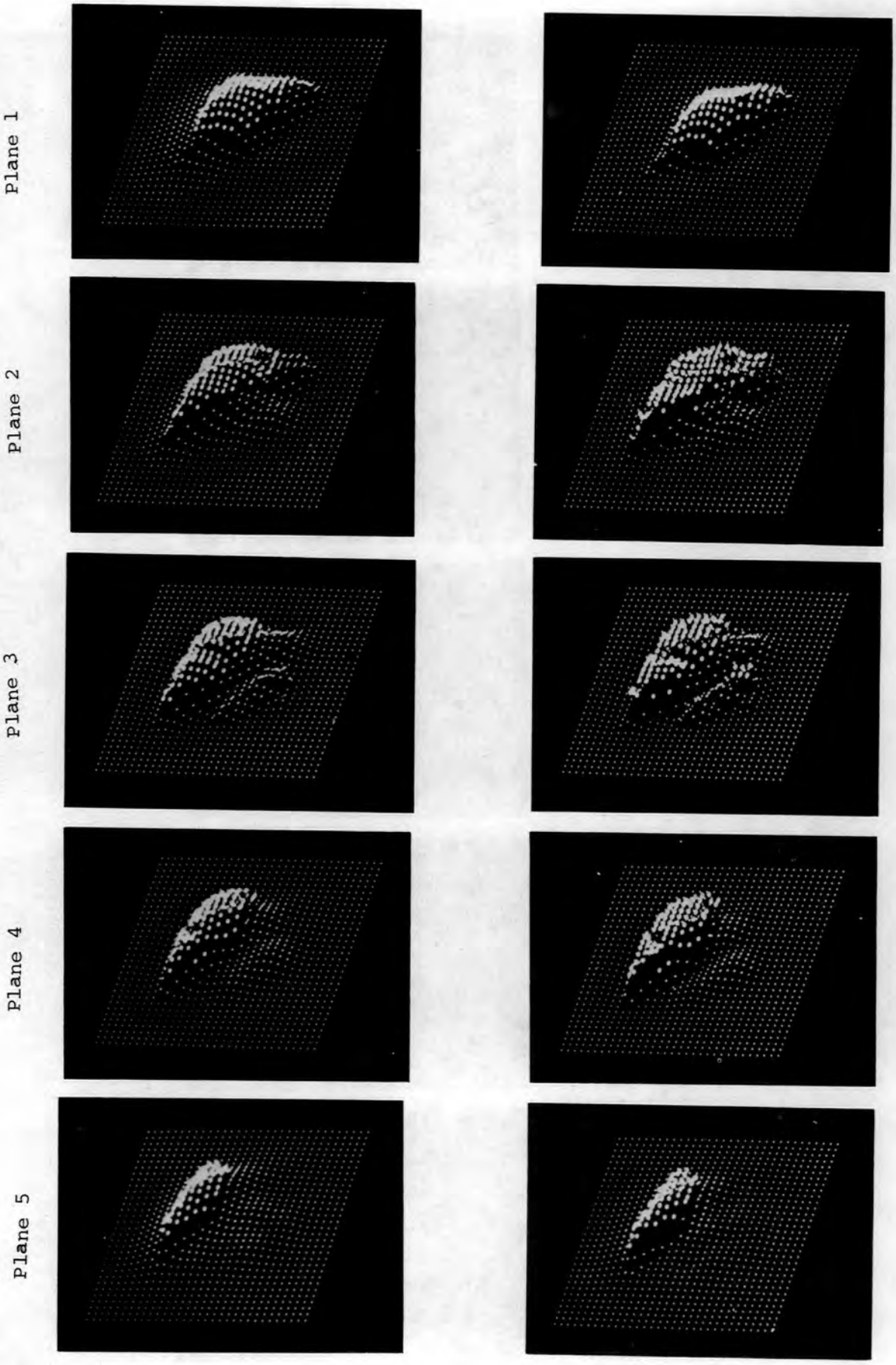
First Iteration Third Iteration

Fig. 12. Restored Images (with noise)



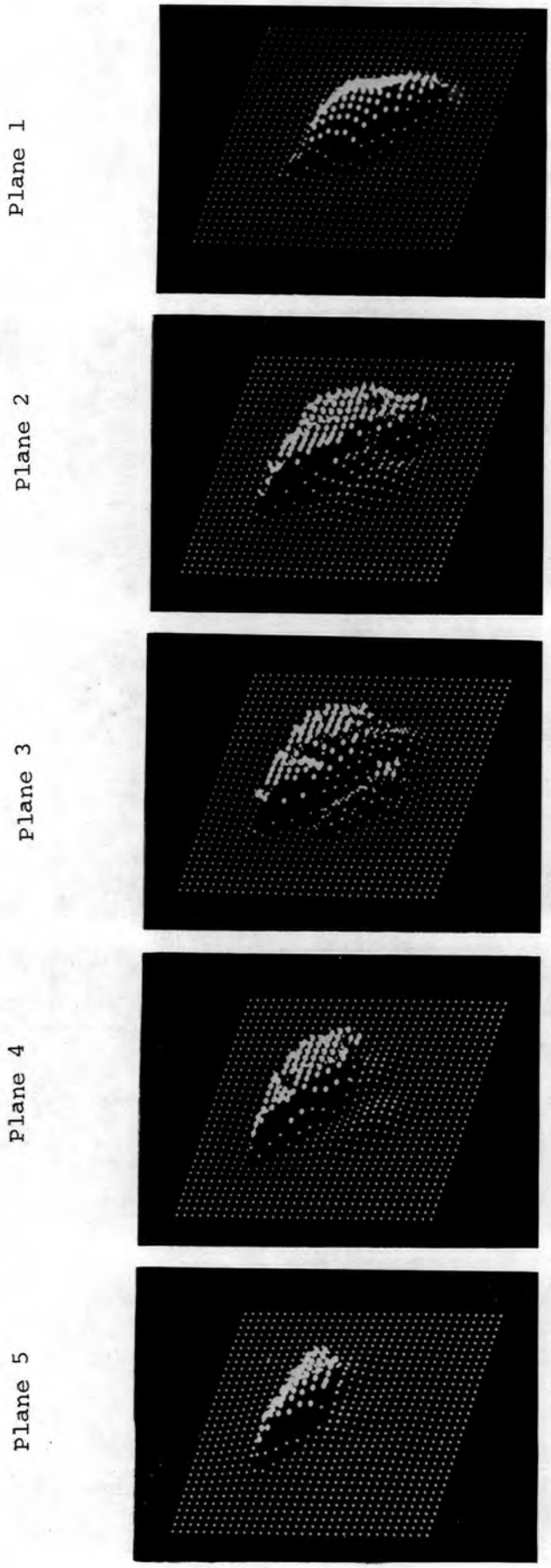
Original Images Observed Images

Fig. 13. Volumetric Display (without noise).



First Iteration 5-th Iteration

Fig. 13. Volumetric Display (without noise)



10-th Iteration
Fig. 13. Volumetric Display (without noise).

of the present method. It is apparent that 3 cold spots can be discerned even with three iterations and that the pancreas can be distinguished from the liver.

Fig. 12 shows the restored images with noise. Here the maxiumum intensity in the original sectional images was assumed to be 400 counts (i.e., 5% statistical noise). Total counts in the whole images are about 170 k counts.

Noise in the observed images is correlated between sectional images, but this effect was not taken into account for simplicity. Covariance between images reduces noise effect in the iteration process, so that this simplification in the simulation leads to larger noise effect in the restored images.

The resultant images in Fig. 12 are essentially the same as Fig. 11a, b and c. Beyond 4 iterations noise increased gradually so that iteration is limited to 3 to 4 times in this example.

In Fig. 13 are shown the restored images without noise in the volumetric display along with the original and the observed images. Note that ghosts of the pancreas appeared in the second and the fourth planes, which might be due to somewhat smoother shape of the pancreas. The liver and cold spots have sharper edges, so that they can be restored easily. This suggests that high frequency components can be restored easily but low frequency components are restored more slowly.

CONCLUSION

With the present iteration method, images of sharp edges can be restored with 2 to 4 iterations but images of smoother edge are more slowly restored. This is due to the characteristics of longitudinal tomography. in which low frequency and D.C. components are lacking. This causes ghosts of the images of smooth edges on the other planes. These ghosts are smoother in shape and may not be significant.

This method will help to distinguish overlapping organs and to find lesions on the periphery of large organs such as liver which are difficult to detect with conventional scintigraphy.

This work was supported in part by a Grant-in-Aid for Cancer Research from Ministry of Health and Welfare.

REFERENCES

1. P. Schmidlin, "Iterative Separation of Sections in Tomographic Scintigrams," Nuclear Medizin, 11, 1 (1972).

2. M. J. Myers, "Computer Processing in Longitudinal Isotope Tomography," 343 in Information Processing in Scintigraphy, C. Raynaud and A. Todd-Pokropek, Editors, 1975.

3. L. T. Chang, B. Macdonald and V. Perez-Mendez, "Axial Tomography and Three Dimensional Image Reconstruction," IEEE Transactions on Nuclear Science, NS-23, No. 1, 568 (1976).

4. G. Chu and K. Tam, "Three-dimensional Imaging in the Positron Camera Using Fourier Techniques," Phys. Med. Biol. 22, 245 (1977).

5. D. K. Faddeev and W. N. Faddeewa, Numerical Calculation Methods of Linear Algebra, Section 2 and Section 3, Government Printing Office for Physical and Mathematical Bibliography, Moscow (1963) (in Russian).

6. T. A. Iinuma and T. Nagai, "Repetitive Correction for a Finite Resolving Power of the Collimator in Scintiscanning," Int. J. Appl. Rad. Isot. 18, 325 (1967).

7. T. A. Iinuma and T. Nagai, "Image Restoration in Radioisotope Imaging Systems," Phys. Med. Biol. 12, 501 (1967).

8. I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, 696, A. Jeffrey, Translation Editor, Academic Press, New York and London, (1965).

DISCUSSION:

The first question directed to Dr. Tomitani regarded the problem of noise. Specifically, was there noise in the original data? Dr. Tomitani answered affirmatively and noted that Poisson noise was added following the blurring operation. He added that this procedure was not strictly rigorous since the noise of each sectional image was correlated with the other sectional images, noting further that the correlated noise of the sectional images tends to reduce the noise in the reconstructed images. Therefore, in the present simulation, noise in the reconstructed images is somewhat exaggerated.

A second question directed to Dr. Tomitani asked if in the generation of the longitudinal tomogram were the same convolution operators used that were used for the generation of the cross-sectional tomogram. Dr. Tomitani's answer was affirmative and noted that in the case of the generation of "the observed images," the events that fall outside the field of view were discarded. He noted further that the number of events that fall within the field of view as a result of the blurring operation would be renormalized so that the total number of events within the field of view is constant. Thus, the total counts of the images are kept constant throughout the reconstruction procedure.

POSITRON EMISSION COMPUTED TOMOGRAPHY USING LARGE AREA DETECTORS*

Frank Atkins, Gerd Muehllehner, and Paul V. Harper
The University of Chicago
The Franklin McLean Memorial Research Institute† and
Searle Radiographics Inc., Product Research
Chicago, Illinois, United States of America

ABSTRACT

The use of large area, position sensitive devices for positron annihilation coincidence detection permit fully 3-dimensional reconstructions of the activity distribution. Since the detectors may remain stationary or be rotated during the study both longitudinal and transverse sections may be imaged. This paper discusses some of the aspects of this imaging geometry, and possible algorithms for these two modes of operation. Also included are some clinical and phantom studies obtained with a positron camera designed around two Searle LFOV cameras.

INTRODUCTION

A number of different positron imaging devices have been developed over the last decade in an attempt to exploit the unique properties of these radionuclides, encouraging new directions of research into 'in vivo' metabolic studies through the use of biologically active compounds labeled with positron emitting tracers.[1] To a large extent, these systems can be classified as section or multi-section devices,[2,3,4] consisting of one or more linear arrays of detectors; and volume imaging devices which consist of position-sensitive, large area detectors.[5,6] It is the later category which will be addressed in this paper. Although most of the discussion will be generally applicable to these systems, the clinical and phantom studies were obtained with the particlar system shown schematically in Fig. 1, which has been described previously in more detail.[7] The detector system consists of two slightly modified, Searle LFOV cameras with 1" thick NaI(Tl) crystals mounted on a rotating gantry. The use of a fast-slow coincidence logic enables the camera electronics to operate at the relatively low count rate due to coincidence events, while the detector electronics operates at the high, noncoincidence count rates. Those events

*This research was supported in part by Contract No. NHLI NO1 HV-52980, from the National Heart and Lung Institute.

†Operated by the University of Chicago for the US Energy Research and Development Administration under Contract EY-76-C-02-0069.

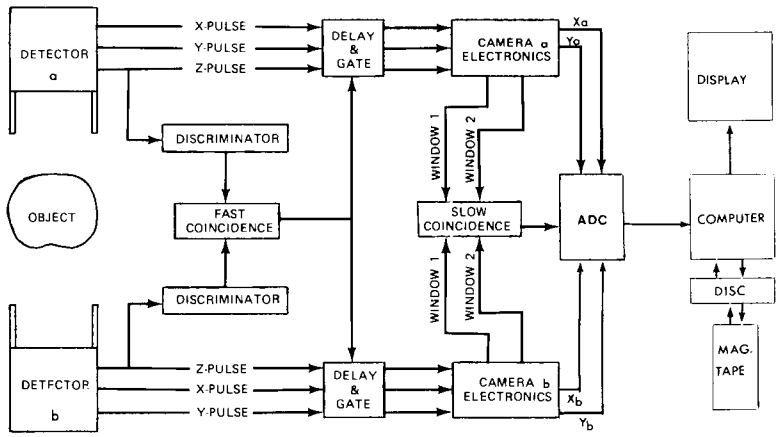


Fig. 1. Schematic diagram of the positron camera system used in this study. The detectors are mounted on a gantry which provides complete rotation and access to the entire body.

which satisfy coincidence and pulse height analysis are passed through an interface to an Interdata 8/32 computer which controls the data acquisition. The positional information of the coincidence events can then be either stored on magnetic tape for subsequent processing or back-projected on-line and stored in core. The final processed images are then viewed on a RAMTEC display.

The advantage of using uncollimated, large area detectors operating in coincidence is that multiple angular projections of the entire object are imaged simultaneously. However, to utilize the information present in this geometry poses a serious computational problem because of the fact that these projections are not restricted to a single plane as is the case with most transmission and emission devices.

Moreover, such a system can be operated in two modes producing either longitudinal or transverse tomographic images. In obtaining longitudinal sections, the detectors remain stationary, while a number of backprojection tomograms are generated. These images may subsequently be processed to improve the contrast by correcting the images for the contributions due to overlying and underlying activity. In the transverse mode, the detectors are rotated 180° during the study with each coincident event being backprojected through a three-dimensional volume, and subsequently filtered. In this paper we will discuss algorithms for both of these imaging modalities and illustrate their performance with some examples of clinical and phantom studies.

For purposes of this discussion we have assumed that the two photons formed in the electron-positron annihilation process are antiparallel, and that this event takes place at the site of emission. The finite range of the positron in tissue and the angular deviation of the photon pair from 180° represent theoretical resolution limits on positron imaging devices, and have been discussed by a number of authors.[8,9]

Transverse Section Reconstruction

Description of Geometry

When two opposed large area detectors are rotated around the object to be imaged, information is collected in such a way that projections through the object are not restricted to lie in parallel planes. Thus, the problem cannot be reduced to a two-dimensional problem as is done in X-ray transmission computed tomography. The advantages of such a system are the high sensitivity and the potential of producing high resolution, quantitatively accurate transverse section images. To illustrate the geometrical relationships, let us first consider the simple case of a point source in the center, halfway between two circular detectors as shown in Fig. 2. With the detectors stationary, the

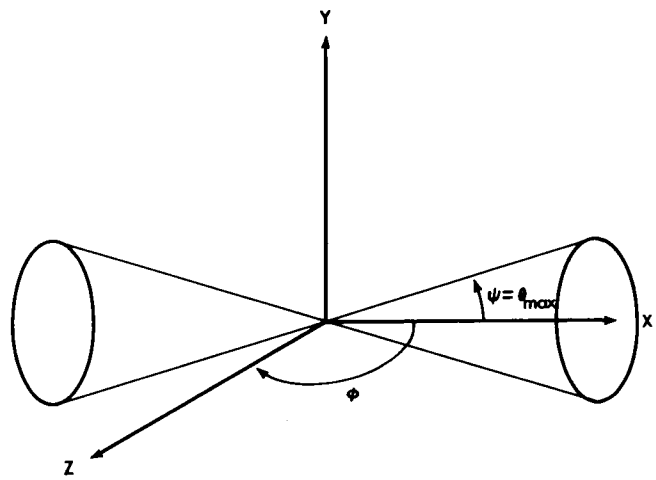


Fig. 2. Coordinate system used to describe the impulse response function in the rotation mode.

point spread function is restricted to lie within the conical volumes shown. Inside this volume the intensity is independent of the rotational angle, ϕ, and varies radially as $1/r^2$. Outside the boundaries of this volume the intensity is zero. As the detectors are rotated, the point spread function becomes symmetrical in ϕ, and the $1/r^2$ fall-off is retained. The functional dependence of the intensity upon ϕ is given by

$$I(\theta) = I(\theta=o)\ (1-\tan^2\theta/\tan^2\psi)^{1/2}\ \tan\psi/\sec^2\theta\ (\tan^2\theta-\tan^2\psi+\sec^2\theta) \text{ for } \theta < \psi \quad (1)$$

where ψ is the maximum acceptance angle subtended by the detectors. The last term in this expression is a result of the decreasing arc length due to circular detectors. $I(\theta)$ is more nearly constant for the case of square detectors. Using Eq. (1) we can write the three-dimensional point spread function as

$$h(r,\theta,\phi) = h(r,\theta) = \text{rect}(\psi-\theta)\cdot(1-\tan^2\theta/\tan\psi)^{1/2}\cdot 1/r^2$$
$$\cdot\ \tan\psi/\sec^2\theta\ (\tan^2\theta-\tan^2\psi+\sec^2\theta) \quad (2)$$

Solution for 4π Symmetry

In order to derive the appropriate inverse filter function we need to obtain the three-dimensional Fourier transform of the point spread function. Let us first solve the simpler case in which ψ is assumed to be $\pi/2$. This

represents infinitely large detectors and results in 4π symmetry. In the case of circular symmetry the Fourier transform is given by[10]

$$H(s, \Theta) = 2\pi \int_0^{\infty}\int_0^{2\pi} h(r,\theta)\, J_0\,(2\pi sr \sin \Theta \sin \theta) \cdot e^{-i2\pi sr \cos \Theta \cos \theta}\, r^2 \sin \theta\, dr\, d\theta \qquad (3)$$

which, in the case of spherical symmetry and with $h(r,\theta) = h(r) = 1/r^2$ reduces to

$$H(s) = 4\pi \int_0^{\infty} \text{sinc}\,(2sr)dr \qquad (4)$$
$$= \pi/s$$

Thus we see that in the case of spherical symmetry the Fourier transform of the point spread function is simply inversely proportional to the radial frequency s (versus ρ in the two-dimensional case). The filter function, therefore, is the familiar ramp filter which can be rolled off at high spatial frequencies to avoid ringing.[11] At the present time we are ignoring the cross-talk between transverse sections and are using a spatial filter restricted to a single plane which is obtained from the ramp filter. We are currently working on obtaining an appropriate solution for the 3-D filter function for the case in which the acceptance angle is restricted to the actual value (approximately 27°).

Method of Implementation

In order to achieve high reconstruction accuracy combined with short reconstruction time we have chosen to backproject each event through a three-dimensional volume and perform the deconvolution operation on the back-projected image. Fig. 3 illustrates the method by which the three-dimensional backprojection is generated. The detectors are rotated as shown through angle ϕ with respect to the Z axis of the fixed coordinate system of the image space.

Each event is defined by camera coordinates Xa, Ya for camera a and Xb, Yb for camera b, and the angle of rotation ϕ. These values can then be converted to object coordinates X', Y', Z' through the transformation:

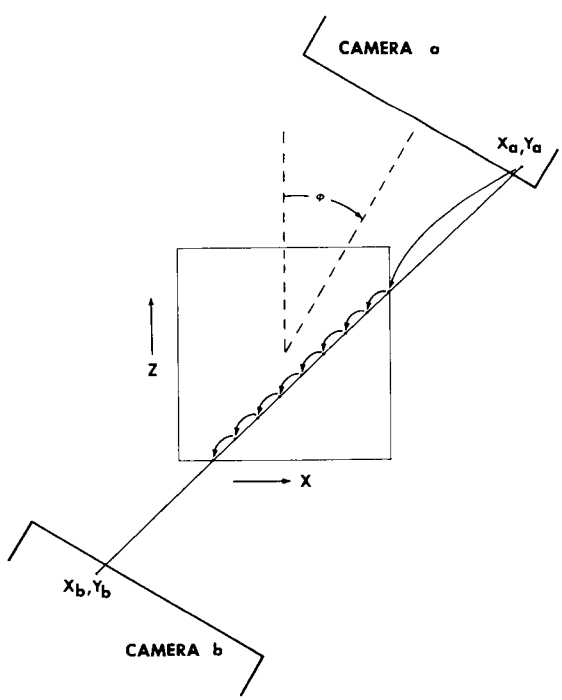


Fig. 3. Ray tracing technique used in generation of the 3-D backprojection images.

$$X'_a = \frac{S}{2} \sin \phi + X_a \cos \phi \tag{6a}$$

$$Y'_a = Y_a \qquad \text{for the event in camera a} \tag{6b}$$

$$Z'_a = \frac{S}{2} \cos \phi - Y_a \sin \phi \tag{6c}$$

and

$$X'_b = - \frac{S}{2} \sin \phi + X_b \cos \phi \tag{7a}$$

$$Y'_b = Y_b \qquad \text{for the event in camera b} \tag{7b}$$

$$Z'_b = \frac{S}{2} \cos \phi - X_b \sin \phi$$

where S is the separation between the detectors.

The straight line connecting these two events represents the back-projected ray. Using a simple ray-tracing technique, points within the object space are calculated at equal distances along the ray. The X, Y, Z coordinates in each step are truncated. The content of the corresponding memory location is then incremented by "one." This technique assures equal projected density along the path and avoids time-consuming calculations involving the volume of overlap between each element and a cylindrical volume surrounding the ray as is sometimes done in two-dimensional backprojection.

Since we wish to perform the backprojection on an event-by-event basis in real time, it is essential to perform these calculations in a minimum amount of time. This has been achieved in the following way. First the X, Y, Z co-ordinates for the initial position are packed into common registers with sufficient precision to avoid image artifacts in the backprojection (1/512 parts of an image element) as shown in Fig. 4. The X, Y, Z increments for a single

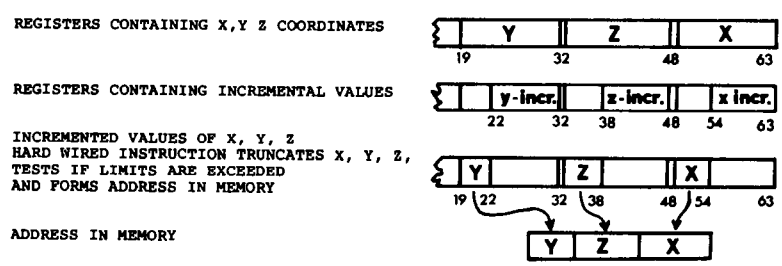


Fig. 4. Computation procedure used in the back-projection to minimize processing time.

step are arranged similarly. Simple register to register addition increments the X, Y, Z coordinates. We have implemented a hardwired instruction which truncates the X, Y, Z coordinates, forms the correct address in memory, tests whether any coordinates has exceeded its upper or lower limit, and terminates the process if the projected ray has reached the edge of the image space. This instruction consists of less than 10 IC chips which requires no further hardware modification to the computer, and can be executed from microcode with an

execution time of 240 nanoseconds.

With the computer connected on-line to the positron camera, the back-projection step can be performed in real time at data rates up to about 6000 counts per second. The filtering operation which is then performed at the conclusion of the study currently takes approximately 10 seconds for each transverse section.

Quantitative Reconstruction

To perform quantitative data analysis does, however, impose certain restrictions; namely a spatially invariant point spread function is required. Therefore, we will discuss factors which, if not compensated for, will introduce spatially dependent variations into the point spread function as well as methods for obtaining a spatially independent point spread function.

A variety of factors such as random counts, count contributions due to scattered gamma rays and effects of attenuation will have a significant effect on the quantitative accuracy of these images. These effects can largely be compensated for and consequently they will not be discussed here since they are present independent of the reconstruction algorithms.

For the point spread function described in Eq. 2 it was assumed that the source was located in the center, halfway between the detectors. If the filter is to be valid for all of the object space, the point spread function must be constant anywhere within that volume.

Consider the X-Z plane shown in Fig. 5, which also shows the detectors in a particular position. For a point source located at the origin the acceptance angle subtended is defined by the useful diameter of either detector. As the point source is displaced along the Z-axis, the solid angle is defined by the useful diameter of that detector which is further from the point source. In a typical case the detector separation is 75 cm and the maximum displacement from the origin is 15 cm for an object which is 30 cm in diameter. For a detector diameter of 38 cm, the angle subtended changes from 27° at the origin to 20° at a distance of 15 cm from the origin along the Y-axis. Restricting the acceptance angle to 20° avoids this variation at some loss in sensitivity. As the source is displaced along the X-axis the acceptance angle subtended changes continuously until, at the edge of the detector, it is reduced to zero. Using the same

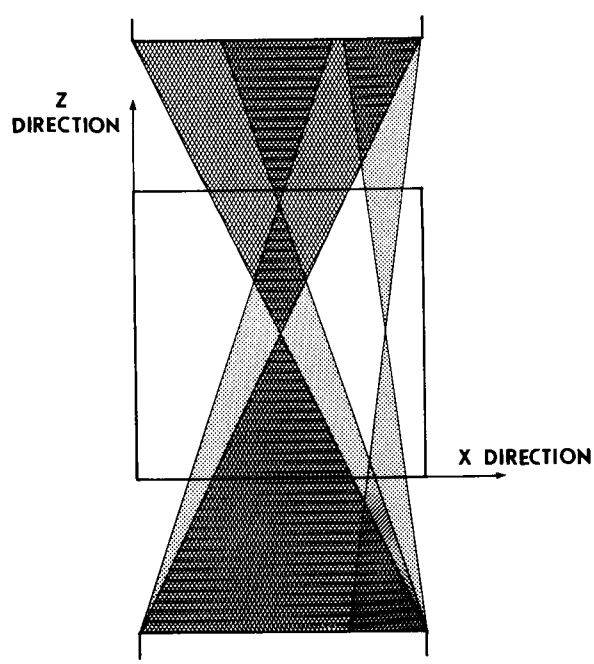


Fig. 5. Variation of sensitivity with source position.

example as above, at a distance of 15 cm from the origin the angle subtended is 6°. This is equivalent to saying that for any particular point in the object, the observation time changes as a function of angle, resulting in angular variations in sensitivity.

The variations of acceptance angle in the X-Z plane combine to produce two effects in the point spread function: (1) for points away from the origin the point spread function is not symmetrical in ϕ and (2) the sensitivity within the plane is not constant. While the latter does not prevent us from performing the filtering operation and can easily be corrected to obtain accurate quantitative data, the former introduces artifacts unless corrected before filtering. If a point source located off-axis is processed with a filter which is symmetrical in ϕ, then the reconstruction is overcompensated in some direction, producing undershooting. This is shown in Fig. 6 for a point source located 13 cm from the center. This effect can be avoided if the intensity of the backprojected ray is made inversely proportional to the observation time for that particular ray during a rotation.

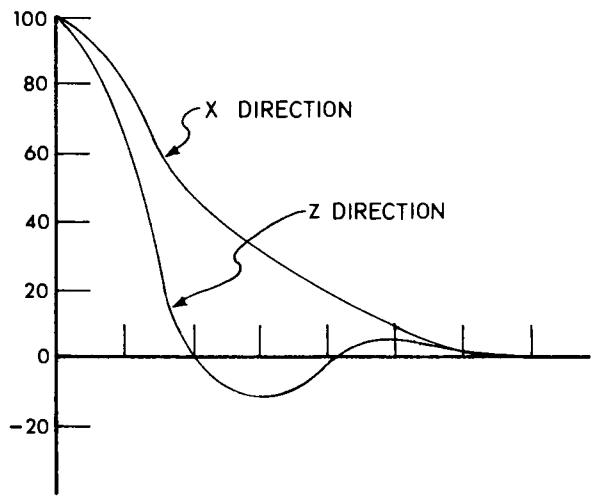


Fig. 6. Nonstationarity in the point spread function produces overcompensation in the Z direction of the reconstruction.

In the above we have only considered variations due to effects in the X-Z plane; as long as square detectors are assumed the acceptance angle in the X-Z plane is constant if the source is displaced along the X-axis. If the source is displaced along the Y-axis, the acceptance angle changes continuously and thus will require a different filter function for each X-Z plane.

Four requirements must thus be met to obtain a spatially invariant point spread function:

1. The detectors must be square or for circular detectors only events falling within an inscribed square must be accepted,
2. The solid angle of acceptance must be restricted to the minimum solid angle encountered in the Z-direction as described above (from a maximum of 27° to 20° for the above example),

3. The backprojected ray must have an intensity which is inversely proportional to the observation time for that particular ray, and
4. A different filter function must be used for each X-Z plane since the acceptance angle, ψ, while constant within an X-Z plane, varies from plane to plane.

Results

Transverse reconstructions have been obtained for a number of different phantoms and in various clinical cases. Fig. 7 shows 4 transverse sections through a cylindrical phantom consisting of a cylindrical source surrounded by scattering material, a cylindrical shell of activity with a diameter of 16 cm and a width of 1 cm. A line source was attached, in addition, as shown in the sketch of Fig. 7 to simulate the sinuses in the brain. The section images are

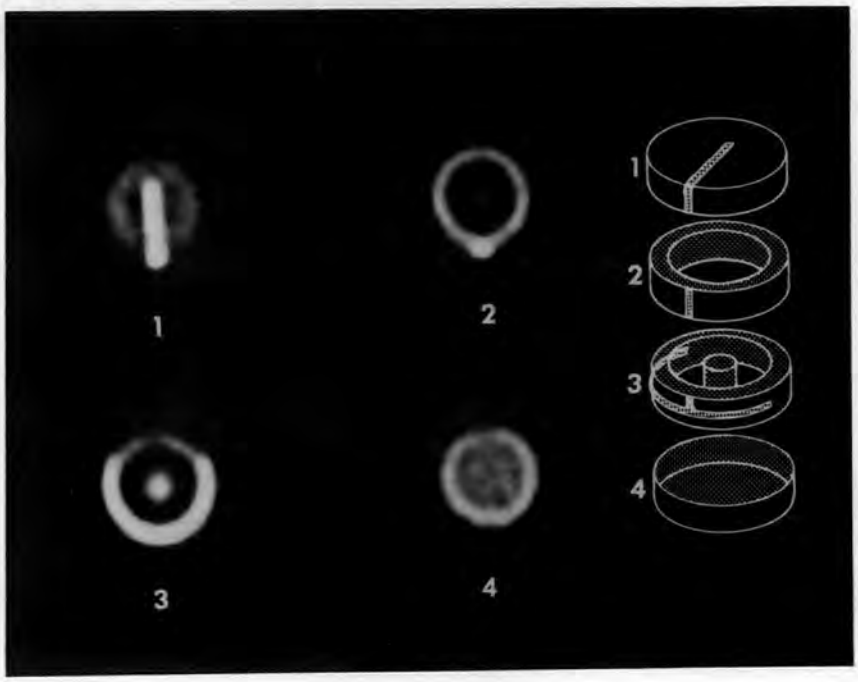


Fig. 7. Study of an idealized brain phantom with a 3 cm diameter lesion in the midline. The drawings on the right indicate the radioisotope distribution for the corresponding four sections. The section thickness is nominally 4 cm, and a total of 1.5 million counts total are contained in the combined images.

seen to correlate very well with the distribution in the phantom, and exhibit the high contrast typical of computed tomography reconstructions. Fig. 8 shows transverse sections through the brain of a normal human subject after inhalation of a gas mixture containing 2 mCi $C^{11}O$. Imaging was started approximately 15 minutes after inhalation, 1.8 million coincident events were collected for all slices with a total imaging time of 20 minutes. Note that the images shown in Figs. 7 and 8 use a two-dimensional filter without restricting the acceptance angle.

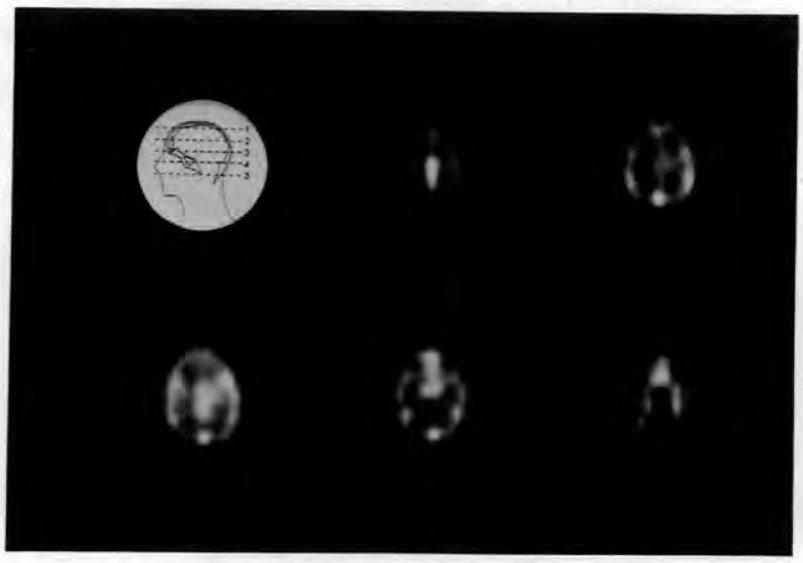


Fig. 8. Transverse sections showing the blood pool in the brain of a normal human subject.

LONGITUDINAL SECTION RECONSTRUCTION

Introduction

In this particular geometry the detector system is assumed to consist of two large area detectors operated in coincidence which view the object from a fixed orientation.

As is the case in the transverse section imaging the large number of coincidence channels which are possible when area detectors are used, make it computationally unfeasible to sort and process the data in this format. The simplest method of pre-processing the data involves the technique of back-projection. The positional coordinates in the detector planes of the coincident photon pair yields a line which passes through the three-dimensional space between the detectors. For any particular event the exact position of the site of emission along this line is unknown. Without any additional information, such as provided by time-of-flight analysis, this point is assumed to be equally likely anywhere along this line. The backprojection image consists of the distribution of the intersection points in an arbitrary plane in the object space which is generally parallel to the detector planes. The technique is simple and in most instances yields acceptable tomographic images, although of lower contrast and quality than those obtained by section imaging devices. It does not, however, yield quantitatively accurate images of the radionuclide density distribution because of the superposition of activity from the entire object. Photons which originate from a source element not on the image plane are recorded as a diffuse distribution with the smoothing becoming more pronounced as distance between the source and image plane increases.

Point Source Blurring Function

The manner in which a point source is defocused as its relative position is changed with respect to the image plane will be described as the blurring function, $b(\vec{\rho},\vec{\rho}_0)$, where the point source is located at $\vec{\rho}_0$ and $\vec{\rho}$ represents an arbitrary point in image space.

Since the annihilation photons are emitted isotropically and anti-parallel, it is sufficient to describe the intensity distribution of just one photon of the pair. Let θ, ξ be the angles which the emitted photon with the coordinate system shown in Fig. 1, for the case in which the angle of rotation, ϕ, equals $\pi/2$. This geometry possesses azimuthal symmetry, and consequently the intensity distribution is only a function of the radial distance. Relative to the Z axis, the number of photons which are emitted into the volume defined between two cones of angles θ and $\theta + d\theta$ is given by

$$dN(\theta) = N_0 \cdot 2\pi \sin\theta \, d\theta \tag{8}$$

where N_0 is the total number of annihilation processes.

The area of the corresponding annulus measured in a plane a distance $\Delta Z = Z - Z_0$ from the source is given by

$$dA(\theta, \Delta Z) = 2\pi (\Delta Z)^2 \cdot \tan\theta \cdot \frac{1}{\cos^2\theta} \, d\theta \tag{9}$$

The intensity is then defined by

$$I(\theta, \Delta Z) \equiv \frac{dN(\theta)}{dA(\theta, \Delta Z)} = \frac{N_0}{(\Delta Z)^2} \cos^3\theta \tag{10}$$

The blurring of a point source as imaged by the positron camera is dependent upon the detector's capability of determining this intensity function in the detector plane. It will be assumed for this analysis that the imaging system has perfect spatial resolution. Thus the blurring function can then be represented by a scalar function

$$b(\vec{\rho}; \vec{\rho}_0) = K \frac{I(\theta, \Delta Z)}{N_0} \omega(\theta) \tag{11}$$

The weighting function $\omega(\theta)$ has been introduced primarily to account for the finite dimensions of the detectors which limit the maximum angle accepted. Because of the digital processing involved with each detected event a weighting factor other than unity or zero could be applied to each photon. Because photons which pass through the object volume at larger angles contain more tomographic information these photons could be given a proportionally greater weight. Such a weighting scheme might improve the quality of the reconstructions but this will not be investigated here.

Formulation of Reconstruction Algorithm

As stated previously the large number of coincidence channels involved with the use of area detectors dictate that the backprojection tomograms represent the first step in the analysis. Implementation of this process can be achieved on-line, as described in a previous section, by keeping the angle of rotation, ϕ, constant.

We proceed from this point by noting that the backprojection tomograms, $t(\vec{\rho})$, of any distributed source, $o(\vec{\rho})$, can be expressed as a three-dimensional integral equation (assuming that attenuation effects have been corrected for by means of a transmission image).

$$t(\vec{\rho}) = \iiint o(\vec{\rho}\,') b(\vec{\rho}\,'; \vec{\rho}) d\vec{\rho}\,' \tag{12}$$

The continuous activity distribution can then be approximated by a finite number

of planes. Corresponding tomograms are then generated for each of these object planes. By restricting the angle which the coincidence ray can make with the detector axis, the blurring functions can be made shift-invariant over the object distribution. As a result the tomogram for any plane becomes a summation of convolutions of the "true" distribution with an appropriate blurring function, $b(\vec{\rho}_i' - \vec{\rho}_j)$. Equation (12) for each of the L tomograms can then be approximated by:

$$t_j \doteq \sum_{i=1}^{L} o_i * b_{ij} \qquad j = 1,\ldots,L \ . \tag{13}$$

The blur function represents the intensity distribution in the jth plane due to a point source located in the ith plane.

Restricting the acceptance angle sacrifices some sensitivity and tomographic effect, but permits a more general analysis of this problem. As a result, Fourier methods can be applied to this reconstruction problem, and some have been discussed by a number of authors.[12,13,14] In this paper we will present a real space, iterative, deconvolution approach to solving Eq. (13). Algorithms of this type have the advantage that the reconstructed values can be constrained locally to have non-negative values and any additional 'a priori' information may be incorporated into the analysis. A major disadvantage, however, is the large amount of computer processing time that is generally required.

One such technique in this general class of processing algorithms is the "Projection Method." This is an iterative technique for solving a system of linear equations,[15] and has been applied to image processing for resolution recovery.[16] Within the context of the longitudinal reconstruction problem the resolution degradation to be compensated for is in the axial direction.

Mathematically the problem can be stated as follows. We are given a degraded tomographic image

$$t(x,y,z) = B[o(x,y,z)] + n(x,y,z) \tag{14}$$

where B is a degrading operator, and $n(x,y,z)$ is a noise term. We wish to find a good estimate for the object, $\hat{o}(x,y,z)$. The approach basically involves attempting to satisfy this equation by iterative means. At each step a new guess solution is determined by the guess at the previous iteration. This is derived from the scalar quantity representing the difference between the observed tomogram at each point and the calculated backprojection value based on the reconstructed images at that stage in the processing. Neglecting the noise term, the digitized version of Eq. (14) can be written as

$$t_{ijk} = \sum_l \sum_m \sum_n b_{i-k,\, j-m,\, k-n}\, \hat{o}_{lmn} \tag{15}$$

and the number of sample points in the reconstruction is taken as the same number in the tomographic images, M, i.e., M = LxNxN, where L is the number of image planes, and N is the dimension of each of these planes. The projection method is best described from a geometrical point of view. We can let $\underline{\hat{o}} = (\hat{o}_{111}, \hat{o}_{112}, \ldots\ldots, o_{l,m,n}) \equiv (\hat{o}_1, \hat{o}_2, \ldots\ldots, \hat{o}_M)$ represent a point in an $\overline{M}$-dimensional space. Then for each (i,j,k), Eq. (15) represents a hyperplane in this space. Let us assume that $\underline{\hat{o}}^\circ = (\hat{o}_1^\circ, \hat{o}_2^\circ, \ldots\ldots \hat{o}_M^\circ)$ represents the initial approximation to the solution. The next approximation $\underline{\hat{o}}^1$ is then obtained by the projection of the point $\underline{\hat{o}}^\circ$ on the hyperplane defined by the point (q,r,s), i.e.,

$$\hat{\underline{o}}' = \hat{\underline{o}}^{\circ} - \frac{(B[\hat{\underline{o}}^{\circ}]-t)_{q,r,s}}{\underline{b} \cdot \underline{b}} \underline{b} \tag{16}$$

where $\underline{b}$ is the row vector of the operator B corresponding to the point (q,r,s), and where $\underline{b} \equiv (b_{q-1,r-1,s-1}, b_{q-1,r-1,s-2}, \cdots, b_{q-L,r-M,S-N})$. This process is repeated for all M points which completes the first cycle of iteration. The order of selection of these M points can be chosen randomly to improve the rate of convergence, as observed by other authors.

It has been shown[17] that the vector sequence $\{\hat{\underline{o}}^{(\circ)}, \hat{\underline{o}}^{(M)}, \hat{\underline{o}}^{(2M)}, \ldots\}$ always converges, and for any value M and coefficients b_{ijk} yields

$$\lim_{p \to \infty} \hat{\underline{o}}^{(pM)} = \tilde{\underline{o}} \tag{17}$$

Furthermore, if Eq. (15) has a unique solution this limit equals that solution; and if infinitely many solutions exist, then $\hat{\underline{o}}$ is the solution that minimizes the quantity $|\tilde{\underline{o}} - \hat{\underline{o}}^{\circ}|^2$. Therefore, for this later situation we can hope to obtain a good solution if we start with a good initial guess. Because the processed images generally maintain similar structural content although with higher contrast as the original backprojection tomograms, then these images represent a good initial estimate.

Dependence of Reconstruction on Imaging Parameters

Because of the limited angular sampling, the quality of the longitudinal reconstructions and the rate of convergence of this algorithm will depend upon the interaction of several parameters; namely, the maximum coincidence angle subtended by the detectors, the spatial extent of the object, and statistical effects. The maximum coincidence angle, θ_{max}, is determined by the dimensions of the detectors and their separation with the constraint that all points of the object be viewed over the same range of angles. Thus in order to achieve an isoplanatic blurring function, the angle θ_{max} is dependent upon the dimensions of the activity distribution as well. A much more serious fundamental limitation of longitudinal reconstruction tomography, is the inability to handle low spatial frequency distributions. The spatial distribution of the activity is, in fact, an important factor in the evaluation of reconstruction algorithms for this mode of imaging.

Several of these parameters have been investigated from an empirical approach using computer simulations of relatively simple objects, namely, point and disk sources confined to a single plane. The rate of convergence of this algorithm is important in view of the amount of computation involved. One measure of the convergence and accuracy is the 'discrepancy,' δ, between the true and reconstructed activity distribution.[18] This has been modified to include the three-dimensional aspect of the problem:

$$\delta_q \equiv \left[\frac{\sum_i \sum_j \sum_l (\hat{o}^q_{ijl} - f_{ijl})^2}{\sum_i \sum_j \sum_l (\hat{o}^{\circ}_{ijl} - f_{ijl})^2} \right]^{1/2} \tag{18}$$

where f represents the true solution.

The value of δ_q represents the ratio of the RMS deviation of the reconstruction from the true value summed over the volume of the reconstruction at

the end of the qth cycle of iteration; relative to the RMS deviation of the initial or backprojection images from the true values.

The maximum coincidence angle accepted has a significant effect on the convergence rate. This is illustrated in Fig. 9, where θ_{max} varies from 10°

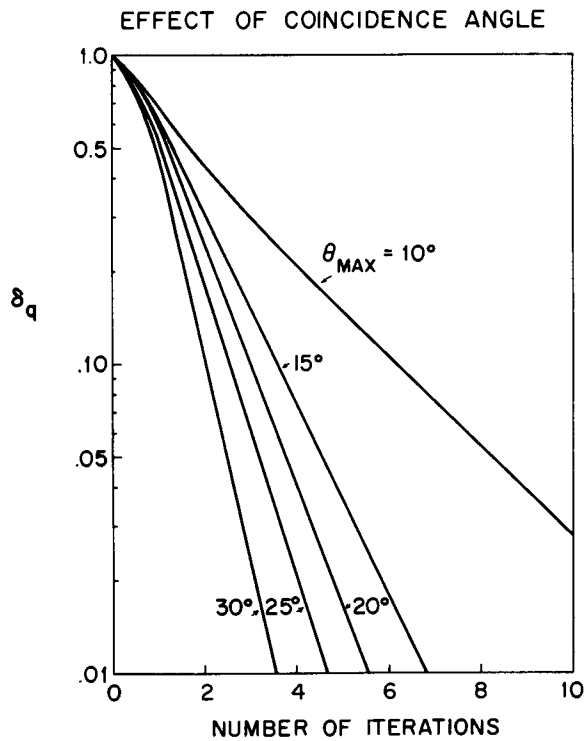


Fig. 9. Rate of convergence of the longitudinal reconstruction algorithm for a set of five tomographic planes separated by 2 cm containing a point source located in one of these planes. This is shown for several values of the coincidence angle of acceptance.

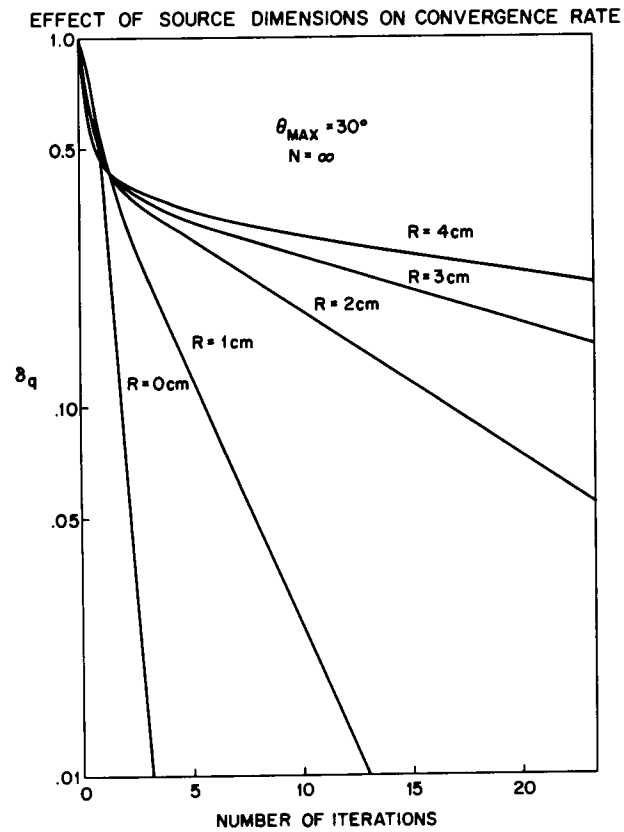


Fig. 10. Effect of the radius of a disk source on the convergence rate. There are five image planes separated by 2 cm used in the simulation. The acceptance angle was assumed to be 30° for all source radii, and the number of protons, N, is infinite.

to 30°. For θ_{max} equal to 30° the discrepancy is less than .01 in 4 iterations, while for 10° this value is about .25.

A more important parameter is related to the dimensions of the source. To study this, a uniform disk source of activity was simulated for the noise-free case. A constant value of θ_{max} equal to 30° was also assumed. The results of these reconstructions are shown in Fig. 10. As the radius of the source increases from a point, the rate of convergence decreases very rapidly. For a uniform source of about the dimensions of the heart it would require a large number of iterations to achieve an accuracy of .01. This is not nearly as serious a limitation as might be expected. An analysis of the spatial redistribution of coincident events, and the influence and statistical fluctuations will illustrate this point.

Quantitative Aspects of the Reconstruction

It is important for any reconstruction algorithm to yield quantitatively accurate processed images. In this situation, this involves a redistribution of the recorded events such that the out-of-focus structures are largely removed from the longitudinal tomograms. Clearly at the start of the processing all the backprojection tomograms contain the same number of events. The objective is to redistribute these events to more accurately represent the true spatial distribution of the origin of the coincident photons. To assess this aspect of the reconstruction, we can define a factor which represents the quantitative dispersion of the axial distribution of activity, Q_ℓ^q,

$$Q_\ell^q \equiv \sum_i \sum_j \hat{o}_{ij\ell}^q \Big/ \sum_i \sum_j \sum_l f_{ijl} \tag{19}$$

The denominator represents the total activity in the object while the numerator is the total activity in the ℓth reconstructed plane at the end of q iterations.

Although the discrepancy for a disk source (diameter - 6.5 cm) is still large after 15 iterations ($\sim$ .20) the relative axial distribution of activity is considerably improved compared to the backprojection images. This effect is illustrated for this example in Fig. 11. It can be seen that the integrated

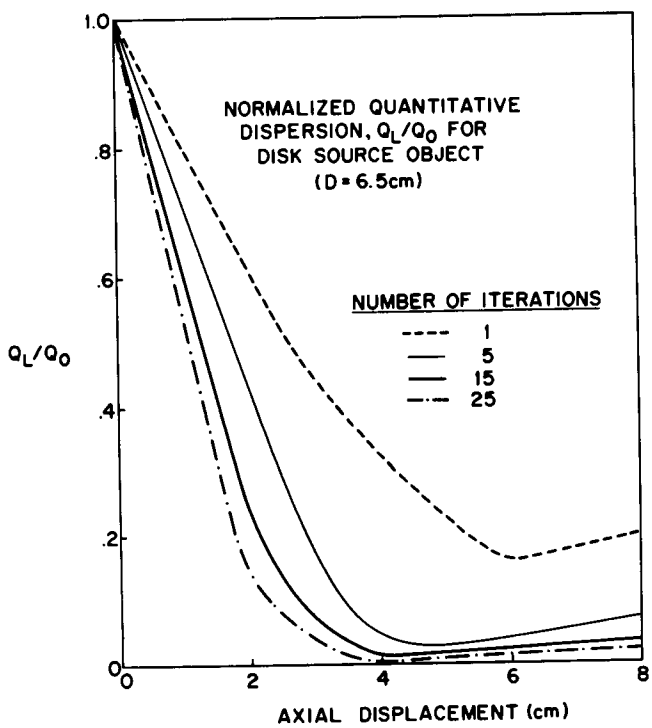


Fig. 11. Dispersion of the total number of events in the set of reconstructed images relative to the source plane. After about 20 iterations there is little improvement in the quantitative aspects of the reconstruction.

activity in the reconstructed set of images relative to the plane containing the source has decreased to a small value (< .05) at an axial separation of 4 cm after only a few iterations.

Effect of Statistical Fluctuations

Up until this point the algorithm has been investigated by noise-free simulations. The effects of statistical noise on the phantom data will be considered next. Because of the correlation of noise between image planes, this

cannot be done simply by generating Poisson distributed counts in each pixel. Instead, the backprojection tomograms were simulated by Monte Carlo procedures in which the angles of emission for an isotropic source were selected randomly. These direction cosines were used to determine the points of intersection through a set of five planes with a 2 cm separation, the first plane of which contains the site of emission. The points of intersection were first determined on a 128 X 128 matrix (1.25 mm) and then reduced to a 32 X 32 size for processing.

Results of these simulations are shown in Figure 12 for the point source reconstruction, where the value of θ_{max} was set to 30°. The discrepancy factor is seen to converge very rapidly to a value proportional to $\sqrt{N}$ for all values of N, which equals the total number of detected events. In order to achieve a value of δ which is less than 0.1 it is necessary to record at least 1000 events. Similar results are observed for the disk sources as well when each resolution element in the object yields on the average the same number of events per pixel as in the point source simulations. The limiting discrepancy values for these conditions are then about the same as the point source values, but the rate of convergence decreases with increasing source dimensions in a manner similar to that shown in Fig. 10.

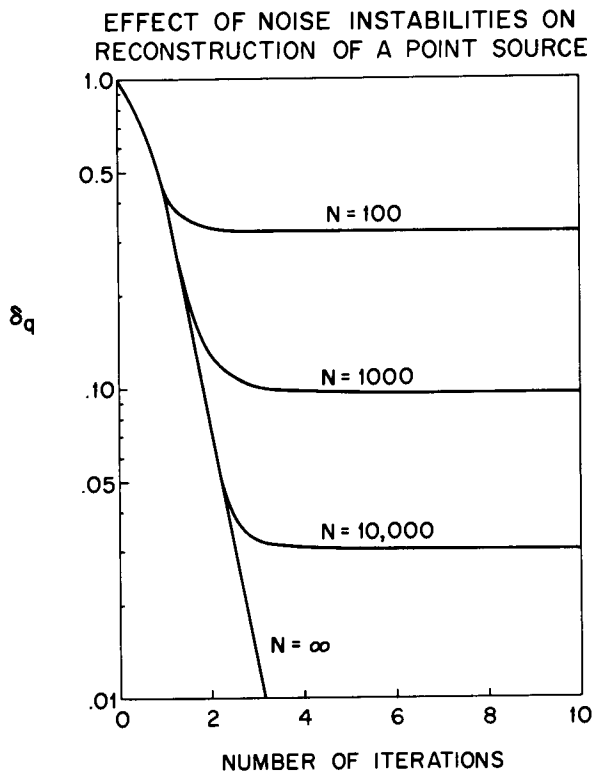


Fig. 12. Statistical aspects of the reconstruction algorithm showing the convergence for a point source with different numbers of detected events.

Clinical and Phantom Studies

To illustrate this algorithm as it is applied to a continuous three-dimensional distribution of activity the following computer simulation was performed. The phantom consisted of a spherical shell of activity surrounding a non-emitting core. The radii of the outer and inner walls were 5 cm and 3.5 cm respectively. This geometry was chosen as an approximation to the myocardium of the left ventricle. Also present in the wall of this phantom is a totally unperfused region with a volume of approximately 5 cm^3. The simulation was again performed using Monte Carlo techniques by randomly selecting the point of emission and the emission angles, and computing the points of intersection of those photons with axial angles < 30° through five (128 X 128) planes with 2 cm

separation. A total of 300,000 photons were recorded. These backprojection images are shown in the left column in Fig. 13. The middle column shows the reconstructions obtained at the end of 20 iterations. The last column shows the backprojection images which would be obtained had only the corresponding 2 cm thick section been imaged separately with the same number of emitted photons. It can be seen that the reconstructions have markedly improved contrast, and more accurately reflect the activity distribution within that volume of the object.

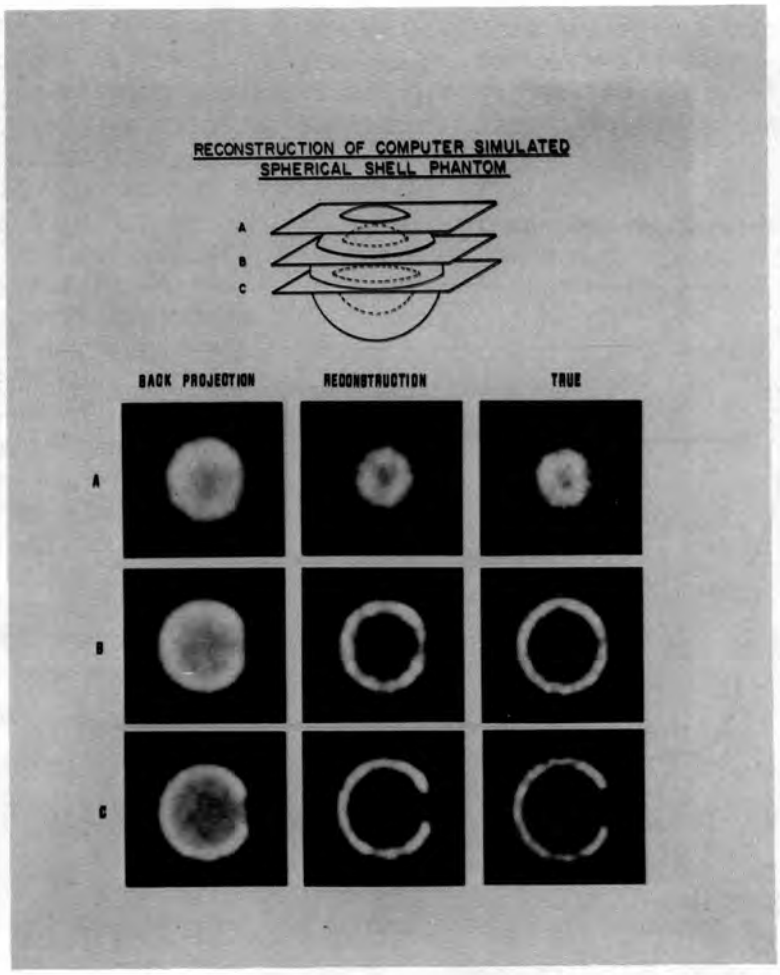


Fig. 13. Computer simulation of a continuous activity distribution. The low contrast backprojection images corresponding to the three planes, A, B, and C are shown in the left column. The reconstructed images in the center show improved contrast and delineation, and correlate with the activity distribution of the corresponding plane shown in the right column.

There is still some extension of the defect into the adjacent plane, and the central region is at a level of about 15% of the activity in the annulus instead of zero.

Examples of clinical studies are shown in the next two figures. Fig. 14 shows the unprocessed and processed longitudinal myocardial tomograms. The emission data ($\sim 10^6$ coincident events) were obtained in about 20 minutes imaging time following the intravenous injection of about 2 mCi $^{13}NH_3$, and allowing approximately 10 minutes to elapse prior to data collection to allow for lung clearance. A set of transmission images were also collected on the patient prior to the injection and were used to correct for attenuation effects.

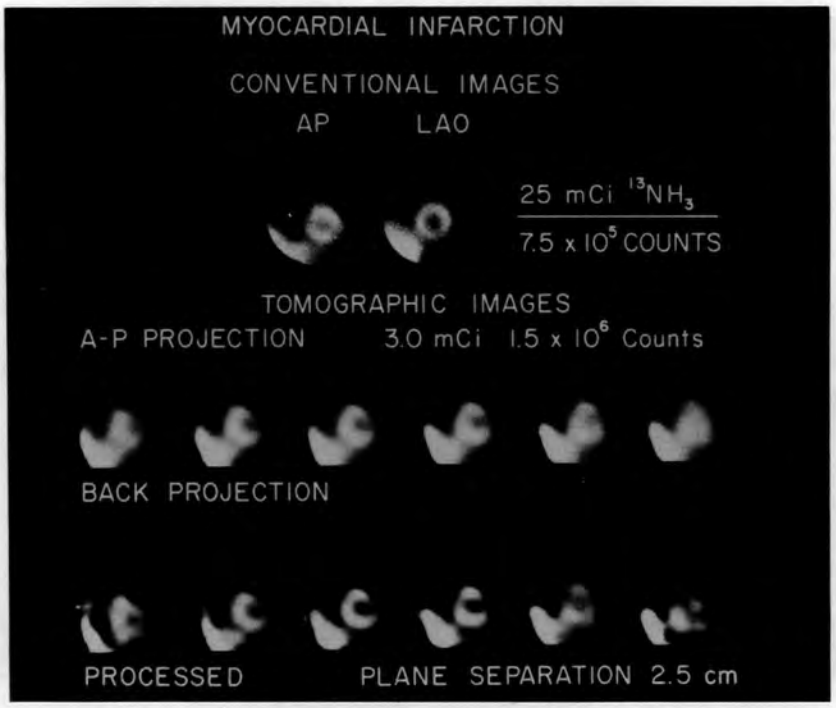


Fig. 14. Example of a clinical study obtained on a patient with a recent inferior postero-lateral myocardial infarction.

Also shown for comparison are the conventional images obtained the same day following a reinjection of $^{13}NH_3$ ($\sim$ 30 mCi) using a high energy tungsten collimator.

Fig. 15 shows longitudinal tomograms of the same object as illustrated in Fig. 8 which depicts the blood pool in the brain obtained after inhalation of a gas mixture containing ^{11}CO.

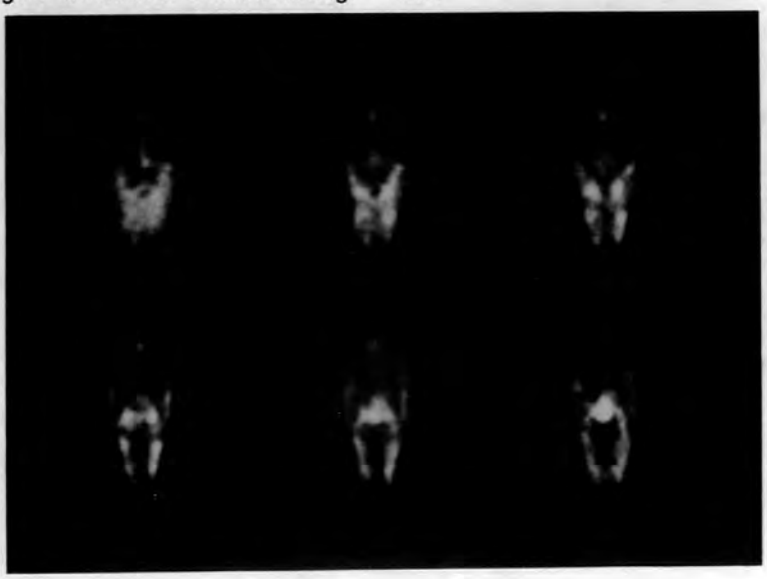


Fig. 15. Series of A-P tomograms of the brain blood pool of a normal human subject using ^{11}CO. The planes are separated by 2.5 cm, and contain a total of 1 million events. The image sequence is posterior to anterior from left to right starting at the top.

CONCLUSIONS

For transverse section reconstructions we have described a technique which uses projections not restricted to lie in parallel planes. The algorithm backprojects each event pair individually which can be performed on-line in real time on a minicomputer. It has been shown that the reconstruction technique

does not lead to serious artefacts in phantom and clinical images; modifications necessary to obtain quantitatively accurate data are also discussed.

Operated in a stationary mode, the positron camera can provide useful data for reconstructing longitudinal tomograms. This is possible despite the limited angular sampling contained in these data. This imaging geometry does, however, have several disadvantages. The limited angular range results in poorer reconstructions for low spatial frequency objects. Also, the input backprojection images have a large distortion of the activity distribution in the axial direction. This can be largely corrected by the processing alorithm, but measurement of defect sizes, e.g., will probably be overestimated. However, processing the backprojection tomograms does yield reasonably accurate results for most object distributions. Furthermore, this mode of operation may be the only choice for situations in which the kinetics of the radiopharmaceutical are fast compared to the imaging time.

ACKNOWLEDGMENTS

We wish to thank Mr. R. Scott for his assistance in the data collection, Mr. J. Dudek for his electronic expertise in modifying and maintaining the system, and Dr. S. Cantez for her support in the clinical studies and evaluation. Also we would like to thank Fran Skozen for preparing this manuscript. Finally, we would like to acknowledge the support of NHLBI Contract No. 1 HV-52980, HL-17648-03, and ERDA Contract EY-76-C-02-0069.

REFERENCES

1. Phelps M., Hoffman E., Kuhl D. Physiologic Tomography (PT): "A New Approach to In Vivo Measure of Metabolism and Physiological Function." IAEA, Symposium on Medical Radionuclide Imaging, Los Angeles, Calif. (1976)

2. Ter-Pogossian M., Phelps M., Hoffman E., et al. "A Positron Emission Transaxial Tomograph for Nuclear Imaging (PETT)." Radiology 114:89-98, 1975.

3. Cho Z., Eriksson L., Chan J. "A Circular Ring Transverse Axial Positron Camera." Workshop on Reconstruction Tomography in Diagnostic Radiology and Nuclear Medicine, Puerto Rico (April 1975).

4. Derenzo S., Zaklad H., Budinger T. "Analytical Study of a High Resolution Positron Ring Detector System for Transaxial Reconstruction Tomography." J. Nucl. Med. 16:1166-1173, 1975.

5. Perez-Mendez V., Kaufman L., Lim C., Price D., et al. "Multiwire Proportional Chambers in Nuclear Medicine: Present Status and Perspectives." J. Nucl. Med. Biol. 3:29-33, 1976.

6. Brownell G., Burnham C. MGH Positron Camera. In Tomographic Imaging in Nuclear Medicine. Soc. of Nucl. Med. 154; 1973.

7. Muehllehner G., Buchin M., Dudek J. "Performance Parameters of a Positron Imaging Camera." IEEE, Trans. Nucl. Sci. NS-23:1, 1976.

8. Phelps M., Hoffman E., Huang S., et al. "Effect of Positron Range in Spatial Resolution." J. Nucl. Med. 16:649-652, 1975.

9. Muehllehner, G. "Resolution Limit of Positron Cameras." J. Nucl. Med. 17:757, 1976.

10. Bracewell R. The Fourier Transform and Its Application, McGraw-Hill, New York, 1965, pp. 252-253.

11. Chesler DA., Riederer, S.J. "Ripple Suppression During Reconstruction in Transverse Tomography." Phys. Med. Biol. 20:632, 1975.

12. Chang L., MacDonald B., Perez-Mendez V. "Axial Tomography and Three-Dimensional Image Reconstruction." IEEE, Trans. Nucl. Sci. NX-23:1, 1976.

13. Muehllehner G., Atkins F., Harper P. "Positron Camera with Longitudinal and Transverse Tomographic Capabilities." IAEA Symposium on Medical Radionuclide Imaging, Los Angeles, Calif. (1976).

14. Chu G., Tam K. "Three-Dimensional Imaging in the Positron Camera Using Fourier Techniques." Phys. Med. Biol. 22: March, 1977.

15. Pizer SM., Numerical Computing and Mathematical Analysis, Science Research Associates, Inc., (1975) pp. 162-169.

16. Huang TS., Barker DA., Berger SP. "Iterative Image Resolution." Applied Optics 14:1165, 1975.

17. Tanaba K. Numer. Math. 17:203, 1971.

18. Oppenheim B. "More Accurate Algorithms for Iterative 3-D Reconstruction." IEEE, NS-21:3, 1974.

DISCUSSION:

Pizer pointed out that the iteration formula used for longitudinal section imaging was based on the method of Kaczmarz, and noted that the formula should be modified slightly because in its stated form it contained an improper vector-by-vector multiplication. He also asked why the bottom slice, which should have been uniform, was non-uniform. Muehllehner responded that the precise reason was not known yet, but thought that the non-uniformity could be due to limited statistics, artifacts due to imperfections in the filter, and cross-talk between slices.

Herman asked whether the iteration formula was actually a form of ART. Atkins felt they were different because the formula required weights representing the contributions from all cells in the volume to the tomographic back-projection image, based on a priori knowledge of the blurring function. Herman maintained that as long as one is solving a system of equations by iterating to reduce the difference between observed and estimated values, this is a form of ART. Rogers felt that the method differs from ART because it involves iteration on a linear combination of projection elements rather than on individual projection elements. Tretiak agreed, noting that this method operates on back-projected data while ART operates on raw projection data.

Todd-Pokropek asked whether any correction for attenuation was being carried out in transverse section imaging. Atkins replied that a correction on each coincidence channel would be very difficult, since the number of coincidence channels can easily approach 10^6, and very few events are detected per channel.

Oppenheim noted that the transverse section images were formed by back-projecting and then filtering, which differs from the usual technique of filtering and then back-projecting. The results looked good, even though a two-dimensional filter was being applied to each plane rather than a three-dimensional filter to the volume. He asked about the shape of the filter, and Atkins replied that it was based on a modified ramp in frequency space with roll-off at high frequencies.

THREE-DIMENSIONAL RECONSTRUCTION BASED ON SEQUENTIAL INFORMATION FROM A SCINTILLATION CAMERA

C. Berche, B. Aubert, A. Bethencourt and R. Di Paola
Institut de Recherche de Radiobiologie Clinique (INSERM U66)
et Institut Gustave-Roussy
Villejuif, France

ABSTRACT

The methods proposed for three-dimensional reconstruction take into account the sequential recording (list mode) of the informations sent by a standard scintillation camera ; the camera is either rotating around the subject, or the subject is rotating in front of the camera. These methods allow to obtain any transverse or oblique section using the entire field of view of the scintillation camera with parallel or fan-beam projections. The first results obtained with phantoms and patients (liver) are presented.

The solution we give to the problem of the three-dimensional (3-D) reconstruction of a radioactive object in an absorbing and scattering medium, is well suited for the standard equipment of a Nuclear Medicine department. In our case, we had at our disposal three scintillation cameras : a Searle Radiographics Pho/Gamma III HP, a Toshiba GCA 401 and a CGR Opticamera.

It is possible to reconstruct transverse sections [1,2,3,4,5,6], or, longitudinal sections [7,8,9]. We decided to use the entire field of view of the gamma-camera, in order to reconstruct any section of the radioactive object. Schmidlin the same choice and implemented different algorithms.[10]

The mathematical techniques detailed later are theoretically equivalent to the classical techniques,but they are different in the explicit use of the notion of continuity. We consider that the collected data represent the ensemble of all the projections of the radioactive object. This is possible with the list mode acquisition of our computer-system[12], recently replaced by an Informatek Simis III System. The x and y coordinates of each event are stored on disk, as well as time information, during the rotation of the radioactive object in front of the camera at a continuous constant speed. This gives a correspondence between the time and the angle of rotation. These data are recorded on a magnetic tape, processed on a Univac 1106 and the resulting pictures displayed off-line on a color TV screen or an electrostatic printer-plotter.

ANALYTICAL RECONSTRUCTION TECHNIQUES

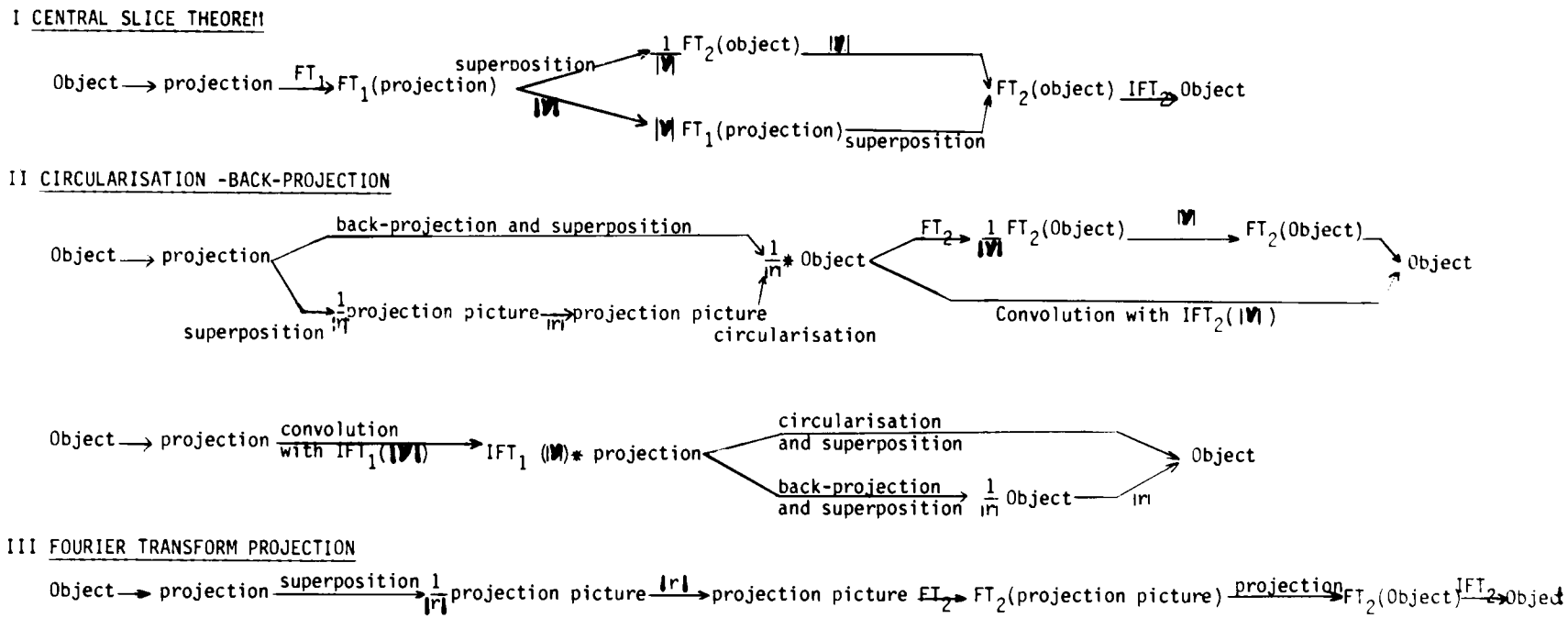


FIG. 1 - Diagram of the fundamental analytical 2-D reconstruction techniques. Each path represented by consecutive arrows corresponds to different useful 2-D reconstruction algorithms.

FUNDAMENTAL TECHNIQUES FOR TRANSVERSE SECTIONS

On a first approximation, the problem of the emission transaxial tomography is similar to the 3-D reconstruction of an object from its projections. These projections can be obtained by integration along any curve. However, we consider only the problem of rectilinear projections, the path of the photons being rectilinear. There are essentially two kinds of projections : the parallel and the particular oblique (or "fan-beam") projections.

Theoretically, if we know the projections of each transverse section of an object we can reconstruct each section separately and also the whole object by piling its reconstructed sections. Here we present only the most fundamental 2-D analytical algorithms for the reconstruction of any transverse section of an object. In practice, these 2-D algorithms must be generalized into 3-D algorithms and modified in order to take into account the limited spatial resolution of the scintillation camera and the noisy data.

This paragraph deals only with transverse sections of a 3-D object f. A point of a transverse section is called f (r,θ) in polar coordinates without mention of the z level of the section considered.

Parallel Projections

All the 2-D analytical techniques presented here are theoretically equivalent. Other authors [13] demonstrate partially the equivalence of these techniques which can be classified into five groups (Fig. 1). We describe only the first three : the Central Slice Theorem techniques[14,15], the Back-Projection or "Circularization" techniques[16,17,18,19,20] and the Fourier Projection Theorem techniques.

Central Slice Theorem

Consider a 2-D transverse section of a 3-D object and its 1-D parallel projection g_θ at any angle θ. Using polar coordinates, this projection can be written

$$g_\theta(r) = \int_0^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\delta(r-r'.\cos(\theta'-\theta)).|r'|.dr'.d\theta' \qquad (1)$$

Here $\delta(r-r'.\cos(\theta'-\theta))$ denotes the Dirac's distribution with unit density on the straight line defined by

$$r-r'.\cos(\theta'-\theta) = 0 \qquad (2)$$

with r and θ fixed and r' and θ' variable.

Then, the 1-D Fourier transform (FT) of g_θ for the frequency ν , is

$$\begin{aligned} FT_1(g_\theta)(\nu) &= \int_{-\infty}^{+\infty} g_\theta(r).\exp(-2i\pi.\nu.r).dr \\ &= \int_0^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\left\{\int_{-\infty}^{+\infty} \delta(r-r'.\cos(\theta'-\theta)).\exp(-2i\pi.\nu.r).dr\right\}.|r'|.dr'.d\theta \\ &= \int_0^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\exp(-2i\pi.\nu.r'.\cos(\theta'-\theta)).|r'|.dr'.d\theta' \end{aligned} \qquad (3)$$

Thus :

$$FT_1(g_\theta)(\nu) = FT_2(f)(\nu,\theta) \qquad (4)$$

In other words, the 1-D FT of any parallel projection at any angle θ of a 2-D function is equal to the central slice at the same angle of the 2-D FT of the 2-D function

$$FT_1(g_\theta)(\nu) = FT_{2,\theta}(f)(\nu) \tag{5}$$

where $FT_{2,\theta}$ is the central slice at the angle θ of FT_2.

Back-Projection or Circularization

The 2-D inverse FT (IFT) of the "extension" of the 1-D FT of g_θ with the notations of the Appendix I, is

$$IFT_2(\tilde{FT}_1(g_\theta)) = \widehat{IFT}_1(FT_1(g_\theta)) = \hat{g}_\theta \tag{6}$$

with : $\tilde{FT}_1(g_\theta)(\nu,\varphi) = FT_1(g_\theta)(\nu).\delta_\theta(\varphi)$ and where o denotes the operation of composition of functions and the symbol $\wedge$ denotes the repetition of the functions considered parallel to themselves. It corresponds to the classical back-projection operation. By superposition (see Appendix I) of the extensions of the $FT_1(g_\theta)$, for all θ, according to the Central Slice Theorem, we get

$$IFT_2(So\tilde{FT}_1(g_\theta)) = IFT_2(1/|\nu|).FT_2(f) = So\hat{g}_\theta \tag{7}$$

where S denotes the superposition operation. Thus

$$(1/|\nu|).FT_2(f) = FT_2(So\hat{g}_\theta) \tag{8}$$

In the real domain, the equivalent equation is

$$(1/|r|)*f = So\hat{g}_\theta \tag{9}$$

where $*$ denotes the convolution and $|r|$ is the module of the radial distance r to the center. Equation (9) corresponds to the classical back-projection technique : each projection is back-projected ($\hat{g}_\theta$) and afterwards all the projections are superimposed ($So\hat{g}_\theta$). The deconvolution can be implemented either using the back-projection, or after the back-projection, in the real or Fourier domains.

The preceding operation can also be interpreted as the integration along a circular path (see Appendix II) in a picture g of which the different g_θ are the central slices. This projection picture $\mathcal{H}'_1(f)$ is defined by

$$g(r,\theta) = \mathcal{H}'_1(f)(r,\theta) \tag{10}$$

where $\mathcal{H}'_1$ denotes the normalized operation of rectilinear projection defined by the equation (1). The circular integration normalized with respect to r and θ is called "circularization" and written :

$$\mathcal{H}'_2(g)(R,\psi) = \int_0^{\pi}\int_{-\infty}^{+\infty} g(r,\theta).|R|.\delta(r^2-r.R.\cos(\psi-\theta)).|r|.dr.d\theta \tag{11}$$

where $|R|$ is a normalization factor of the circularization. Since the superposition introduces the weighting factor $1/|R|$, we get

$$So\hat{g}_\theta = \frac{1}{|R|}.\mathcal{H}'_2(g) = \mathcal{H}_2(g) \tag{12}$$

where $\mathcal{H}_2$ denotes the circularization operator with density $1/|R|$, so that $\mathcal{H}_2(g)$ can be written

$$\mathcal{H}_2(g)(R,\psi) = \int_0^{\pi}\int_{-\infty}^{+\infty} g(r,\theta).\delta(r^2-r.R.\cos(\psi-\theta)).|r|.dr.d\theta \tag{13}$$

Then, from equation (8), it follows that

$$\frac{1}{|r|} \cdot FT_2(f) = FT_2 \circ \mathcal{H}_2 \circ \mathcal{H}'_1(f) \tag{14}$$

The equivalent equation in the real domain is

$$\frac{1}{|r|} * f = \mathcal{H}_2 \circ \mathcal{H}'_1(f) = \mathcal{H}_2(g) \tag{15}$$

This equation shows that each projection can be reconstructed by the formation of the projections picture g and the circularization of this picture.

Fourier Projection Theorem

The 1-D FT of any projection g_θ of f, can be considered as the projection of a 2-D function G. Then the ensemble of all the $FT_1(g_\theta)$ is the ensemble of all the projections of the function G. In this case, it is possible to use the Central Slice Theorem which is again true with the IFT for the kind of functions considered. Then, the ensemble of the g_θ (i.e.g) is the 2-D IFT of G

$$g = IFT_2(G) \tag{16}$$

or

$$G = FT_2(g) \tag{17}$$

and

$$FT_1(g_\theta) = \mathcal{H}'_{1,\theta}(FT_2(g)) \tag{18}$$

where $\mathcal{H}'_{1,\theta}$ is the central profile of $\mathcal{H}'_1$ at angle θ. The 1-D FT of a central slice of g is equal to the projection at the same angle of the 2-D FT of g ; this is analogous to the reciprocal theorem of the Central Slice Theorem. By superimposition of the extensions of all the $FT_1(g_\theta)$, we obtain

$$SoF\tilde{T}_1(g_\theta) = So\tilde{\mathcal{H}}'_{1,\theta}(FT_2(g)) = \frac{1}{|r|} \mathcal{H}'_1 \circ FT_2(g) \tag{19}$$

with

$$SoF\tilde{T}_1(g_\theta) = \frac{1}{|r|} \cdot FT_2(f) \tag{20}$$

It follows that

$$FT_2(f) = \mathcal{H}'_1 \cdot FT_2(g) \tag{21}$$

or

$$FT_2 = \mathcal{H}'_1 \circ FT_2 \circ \mathcal{H}'_1 \tag{22}$$

This equation, considered as fundamental, shows that the inverse operation of the parallel projection in the real domain is the similar projection in the Fourier domain.

Other Analytical Techniques

The technique described by Radon[21] is theoretically equivalent to the preceding algorithms. Others techniques [22,23], difficult to use in practice, will not be considered here.

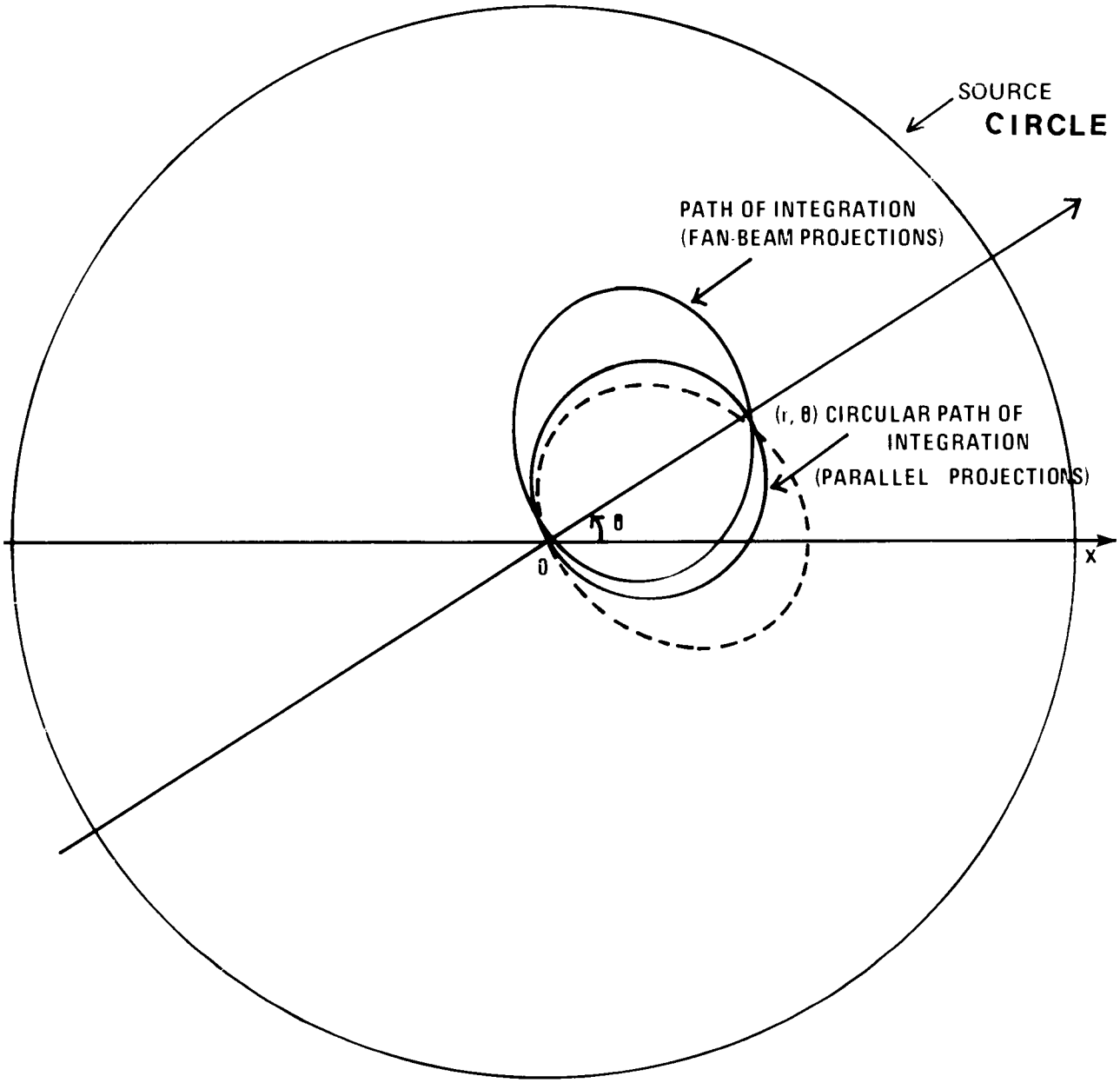


FIG. 2 - Fan-beam projections. One can see the path of integration for the fan-beam reconstruction of the point (r,θ) and the corresponding circular path of integration for the parallel reconstruction of the same point.

Fan-Beam Projections

In this section, oblique rectilinear projections, convergent or divergent from the point source, are considered. Techniques similar to those described for the parallel projections can be implemented, but, due to the complexity of the equations, their use is not easy. Figure 2 illustrates the integration path for these fan-beam projections which can be compared to the circular path described for the parallel projections.

It seems to be more interesting to transform the data in order to come back to the parallel case [24,25].

Consider the general case of the fan-beam projections with the projection of the point source at the distance r_o of the projection of the rotation center. Then, any projection g_θ at the distance D of the point source can be written in polar coordinates.

$$g_\theta(r) = g(r,\theta) = \int_0^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\delta\left[\frac{D.(r-r'.\cos(\theta'-\theta))-(r-r_o).r'.\sin(\theta'-\theta)}{\sqrt{D^2+(r-r_o)^2}}\right].r'.dr'.d\theta' \tag{23}$$

where the numerator of the expression within brackets is the equation of the oblique line passing through (r,θ) and the point source

$$D.(r-r'.\cos(\theta'-\theta))-(r-r_o).r'.\sin(\theta'-\theta) = 0 \tag{24}$$

The denominator $\sqrt{D^2+(r-r_o)^2}$ is introduced to normalize δ to unit density. Then, if we define

$$\cos\varphi' = \frac{D}{\sqrt{D^2+(r-r_o)^2}} \tag{25}$$

$$\sin\varphi' = \frac{r-r_o}{\sqrt{D^2+(r-r_o)^2}} \tag{26}$$

equation (24) gives

$$r.\cos\varphi'-r'.\cos(\varphi'+\theta-\theta') = 0 \tag{27}$$

Since

$$\rho = r.\cos\varphi' \tag{28}$$

$$\varphi = \varphi'+\theta \tag{29}$$

the substitution in (27) yields

$$\rho-r'.\cos(\varphi-\theta') = 0 \tag{30}$$

which is the equation of a straight line passing through (ρ,φ) where the variables are (r',θ').

Therefore, equation (23) can be written

$$g(r,\theta) = \int_0^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\delta(\rho-r'.\cos(\varphi-\theta')).|r'|.dr'.d\theta' = \bar{g}(\rho,\varphi) \tag{31}$$

The transformed projection picture $\bar{g}$ is a function of (ρ,φ) which results from the transformation of g defined by the equations (25), (26), (28) and (29), and lies within the point source circle. Due to equation (31), all the preceding parallel algorithms can be used to fan-beam projections in replacing g by $\bar{g}$ or

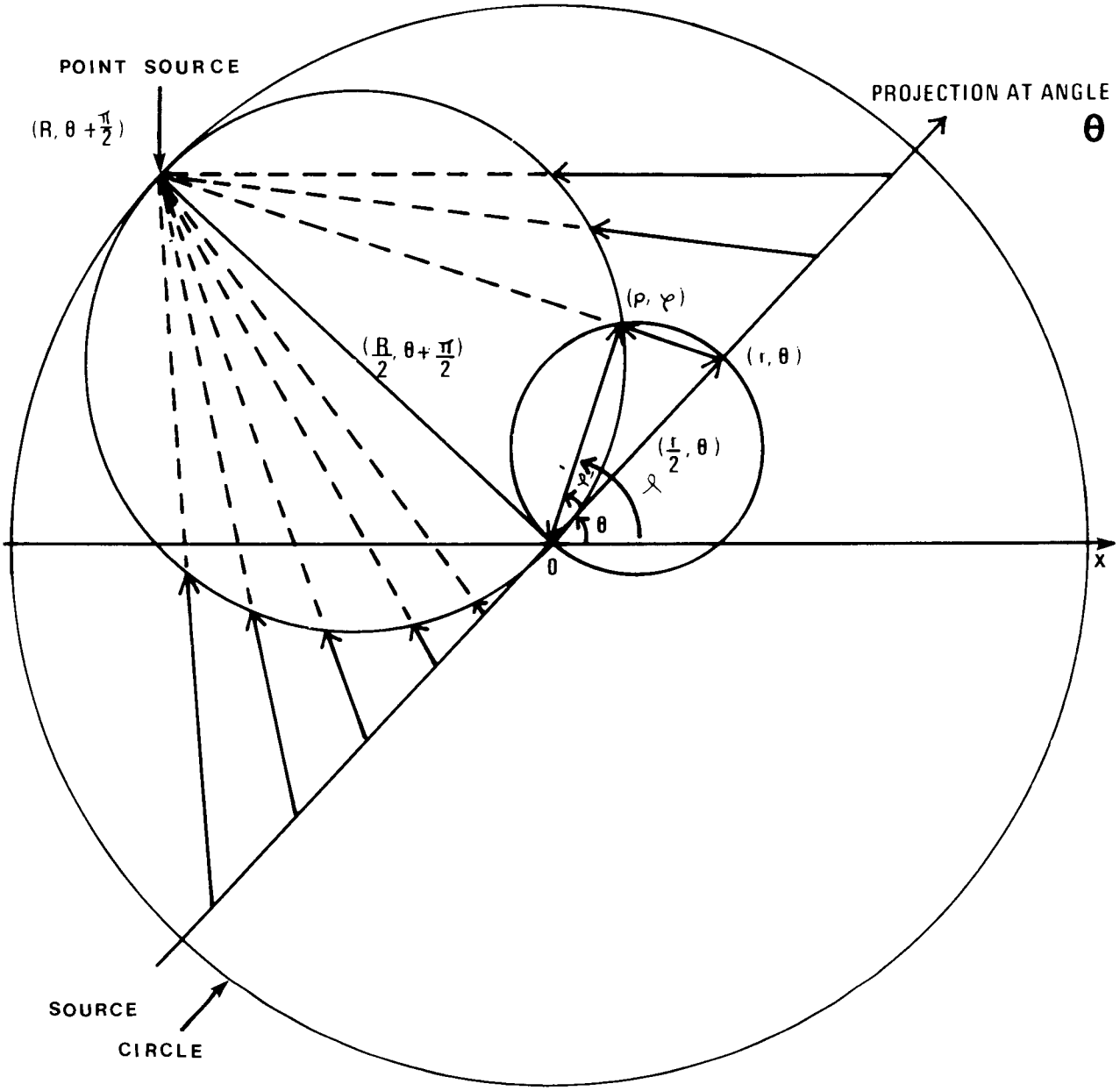


FIG. 3 - Transformation of fan-beam projections into parallel projections. The arrows show the circular distorsion of each fan-beam projection.

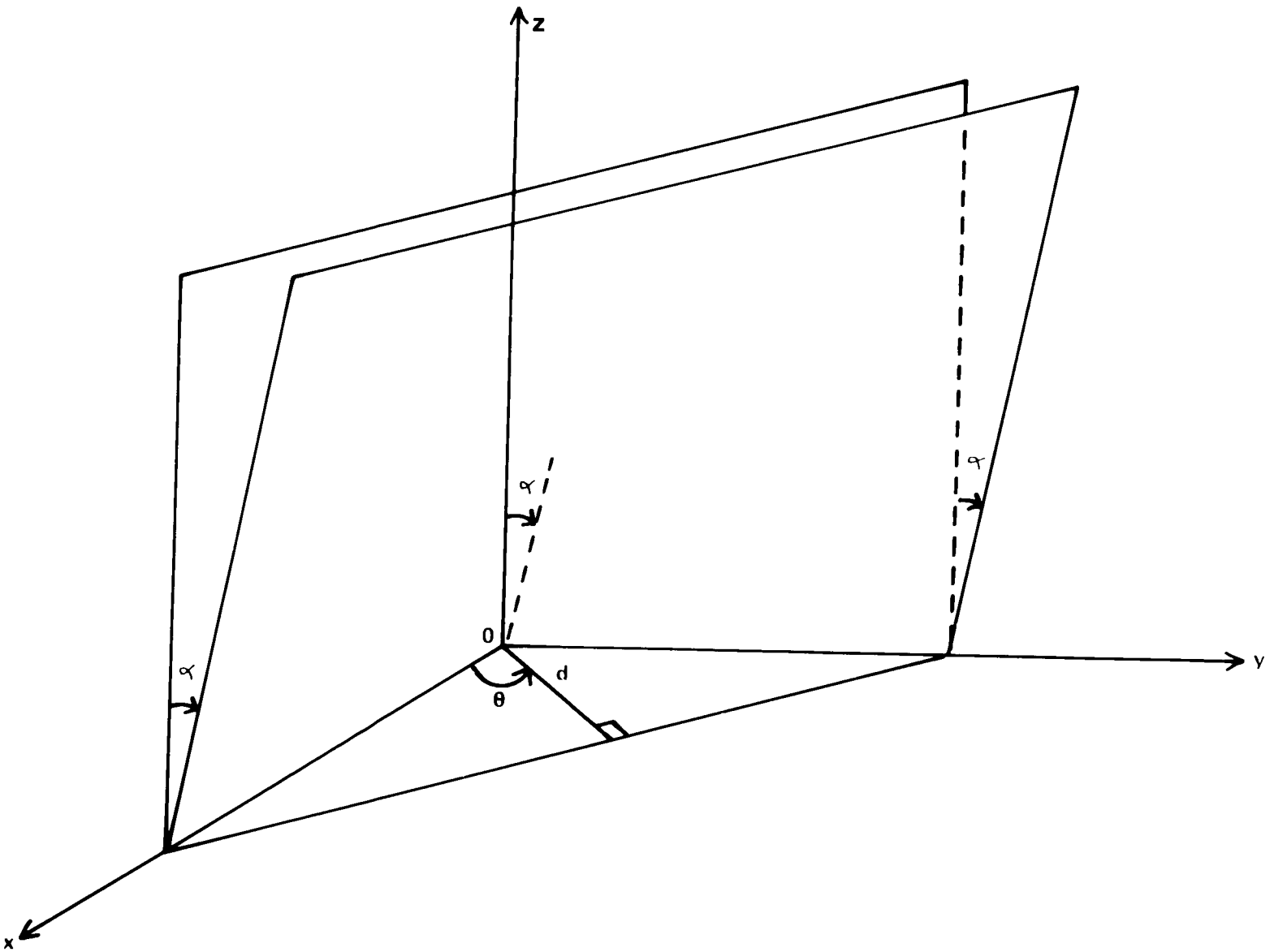
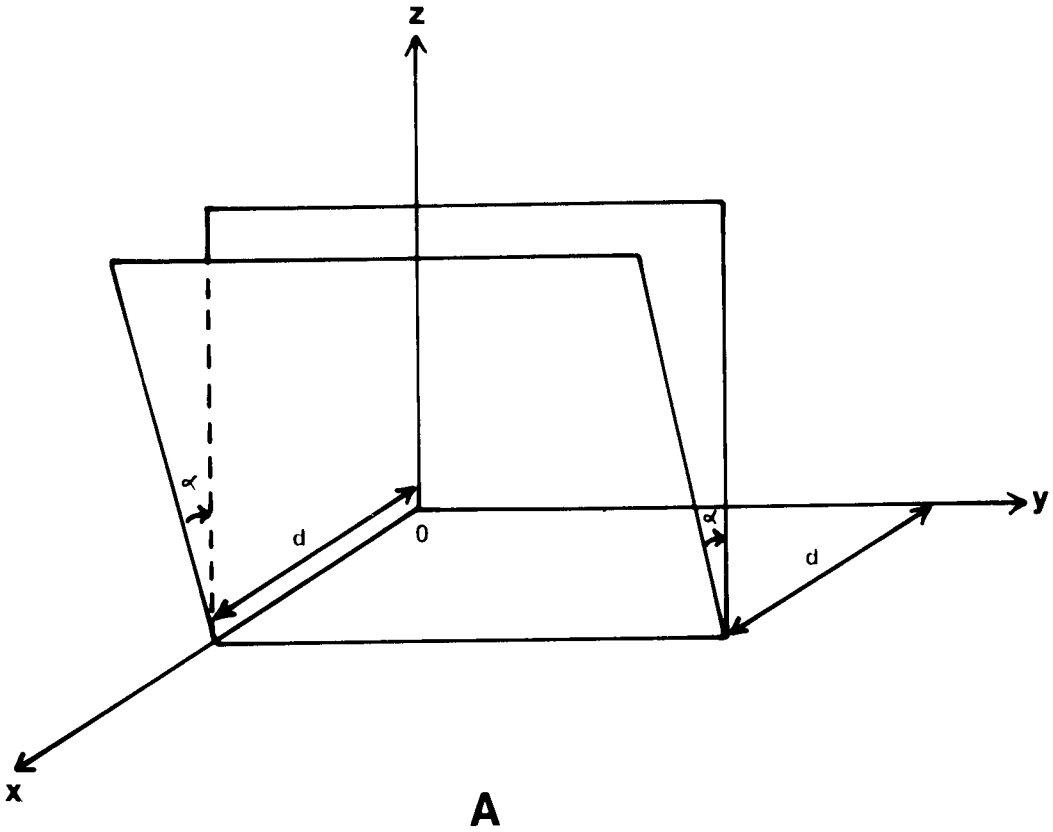


FIG. 4 - Oblique sections in the real domain are defined by three parameters : the angle of obliquity α, the rotation angle θ, and the distance d between the rotation axis and the intersection of the oblique plane of the reference transverse plane (xOy).



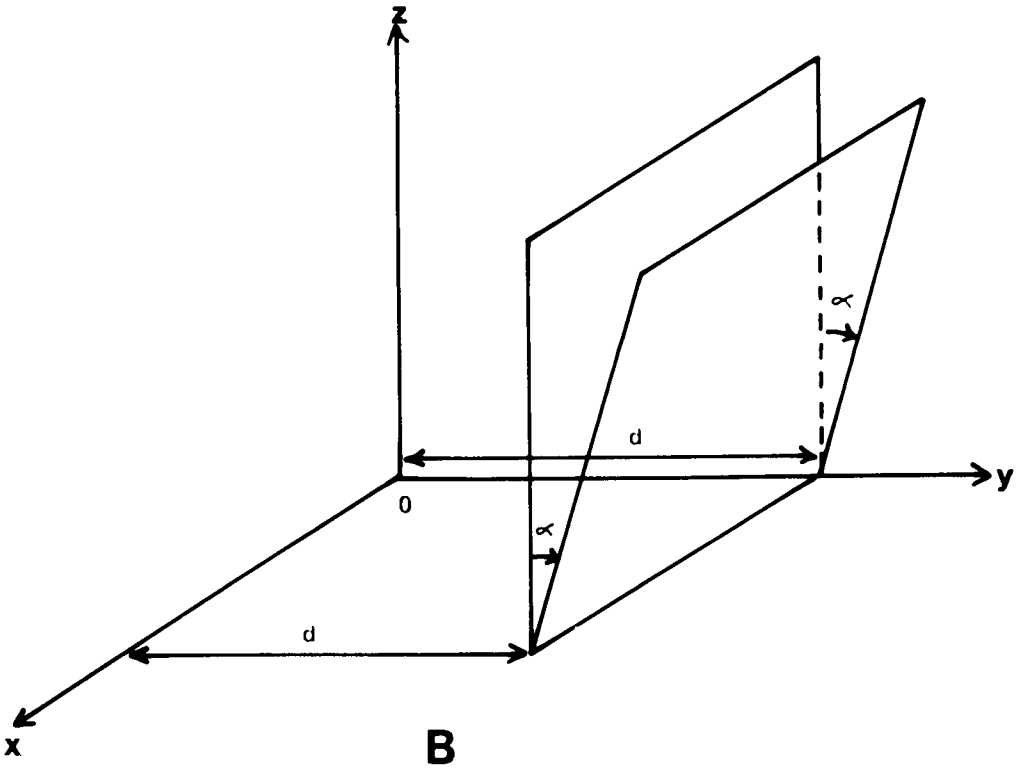


FIG. 5 - Frontal (B) and sagittal (A) sections are parallel to the rotation axis (Oz) and only defined by their distance d to this axis.

$\bar{g}_\varphi$. To perform this transformation, each projection at the center of rotation is distorted circularly : the resulting circle passes through the center of rotation, is tangent to the point source circle and its diameter is equal to the radius d of the point source circle (Fig. 3). This radius d is only useful to the knowledge of the final dilation or contraction factor (D/d) for the reconstructed picture.

This transformation allows to avoid the problem of incomplete projections[26] when the field of the scintillation camera is small or the organ examined is large, using a divergent collimator. With a large field scintillation camera, this transformation is mainly useful to convergent collimators. With usual convergent or divergent collimators, it is necessary to modify the preceding equations to take into account their conical geometry.

PLANE OBLIQUE SECTIONS

Using the techniques previously described, we can obtain theoretically the 3-D object by direct reconstruction. Thus we can also obtain any section of this object. Here, we only consider the plane oblique sections which are clinically interesting.

Consider a 3-D object f(x,y,z). Let (Oz) be the rotation axis and (Ox) the axis of the initial data collection, then an oblique section can be defined by the following three parameters (Fig. 4) : the angle of obliquity α, the rotation angle θ and the distance d between the rotation axis and the intersection of the oblique plane and of the reference transverse plane (xOy).

Two particular cases are interesting : frontal and sagittal sections (Fig. 5). Usually, a frontal section is a section parallel to an anterior or posterior face of the human body and a sagittal section is a section parallel to a right or left profile of the human body. By extension, if the beginning of the data collection corresponds to an anterior or posterior face, one obtains an oblique frontal section when the angle θ is equal to zero (Fig. 5-B). An oblique sagittal section corresponds to an angle θ equal to 90 degrees (Fig. 5-A). These sections are then only defined by two parameters : the obliquity α and the distance d. If α is equal to zero, one obtains the conventional frontal and sagittal sections which are only defined by their distance to the rotation axis.

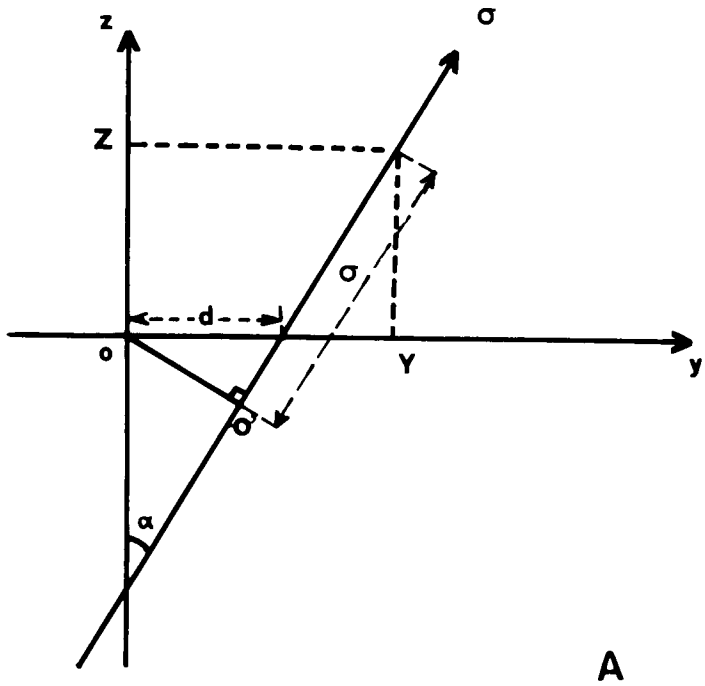
Fundamental Equations

To simplify, one can, for example, consider the case of any oblique frontal section with an obliquity α and at a distance d of the rotation axis (Oz) in the reference plane (xOy). All other cases can be deduced from this case by rotation around the rotation axis. This oblique frontal section is defined by

$$y = d + z.\mathrm{tg}\,\alpha \tag{32}$$

for all x and with $\alpha \neq \pi/2 + k.\pi (k = 0, \pm 1, \pm 2, \ldots)$.

One can define a new system of reference axes from the section of interest. For example, the center of the new axes is the point O', projection of the reference center O on the plane section. The new axes are (O'x) (since it is a frontal section) and (O'σ) (Fig. 6-A). The variable σ depends on y and z according to



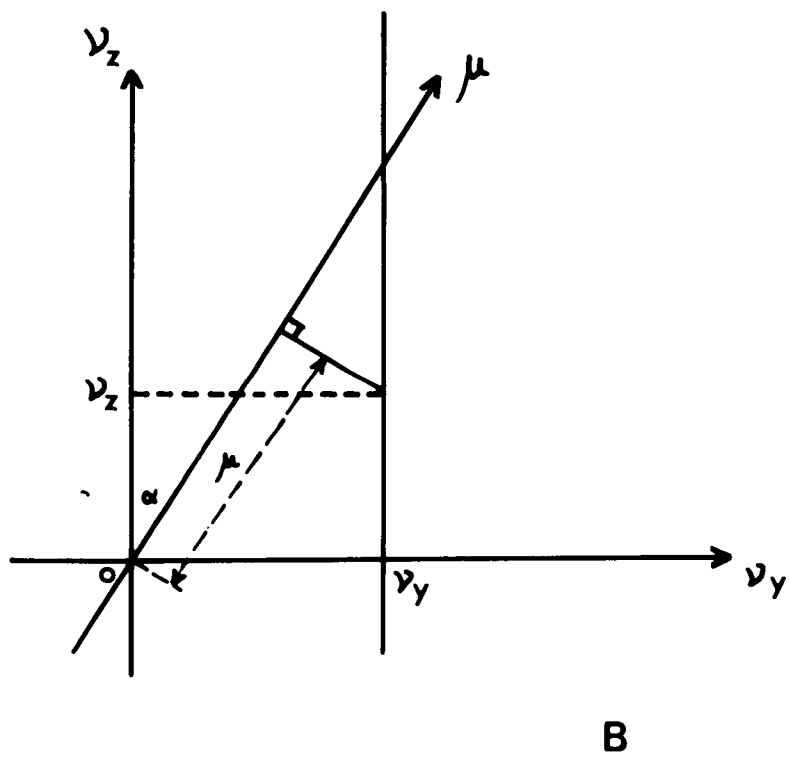


FIG. 6 - Frontal oblique sections : definition of a new real axis (O'σ) (diagram A) and the corresponding new axis (Oμ) in the Fourier domain (diagram B).

$$\sigma = y.\sin\alpha + z.\cos\alpha \tag{33}$$

or, using (32)

$$\sigma = d.\sin\alpha + z/\cos\alpha \tag{34}$$

Then, z and y can be expressed as functions of σ, x and d

$$z = (\sigma - d.\sin\alpha).\cos\alpha \tag{35}$$

$$y = d + (\sigma - d.\sin\alpha).\sin\alpha \tag{36}$$

Consider the 3-D FT $A(\nu_x,\nu_y,\nu_z)$ of a 3-D object f. The 3-D IFT of A at the point (X,Y,Z) is

$$f(X,Y,Z) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} A(\nu_x,\nu_y,\nu_z).\exp(2i\pi.(\nu_x.X+\nu_y.Y+\nu_z.Z)).d\nu_x.d\nu_y.d\nu_z \tag{37}$$

Using (35) and (36), we get

$$\nu_y.Y+\nu_z.Z = \nu_y.d+(\sigma-d.\sin\alpha).(\nu_y.\sin\alpha+\nu_z.\cos\alpha) \tag{38}$$

Consider the new frequency variable μ, defined by (Fig. 6-B)

$$\mu = \nu_y.\sin\alpha+\nu_z.\cos\alpha \tag{39}$$

Then equation (37) can be written

$$f(X,Y,Z) = f_d(X,\sigma) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} A(\nu_x,\nu_y,\mu).\exp(2i\pi.(\nu_x.X+\nu_y.d+(\sigma-d.\sin\alpha).\mu) \cdot |J|.d\nu_x.d\nu_y.d\mu \tag{40}$$

where J denotes the Jacobian determinant obtained when the variables (ν_x,ν_y,ν_z) are changed into (ν_x,ν_y,μ) so that $|J| = 1/|\cos\alpha|$. Thus

$$f_d(X,\sigma) = \frac{1}{|\cos\alpha|}.\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} A(\nu_x,\nu_y,\mu).\exp(2i\pi.\nu_y .d).d\nu_y\right] . \exp(2i\pi.(\nu_x.X+(\sigma-d.\sin\alpha).\mu)).d\nu_x.d\mu$$

$$= \frac{1}{|\cos\alpha|}.\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} IFT_{1,\nu_y}(A)(\nu_x,\mu;d).\exp(2i\pi.(\nu_x.X+(\sigma-d.\sin\alpha)).d\nu_x.d\mu \tag{41}$$

where $IFT_{1,\nu_y}(A)(\nu_x,\mu;d)$ denotes the 1-D Fourier transform of A with respect to ν_y, at the point (ν_x,μ) and depending on the parameter d. This can be written

$$f_d(X,\sigma) = \frac{1}{|\cos\alpha|}.IFT_2\left[IFT_{1,\nu_y}(A)(\nu_x,\mu;d)\right](x,\sigma-d.\sin\alpha). \tag{42}$$

where the 2-D IFT is calculated over the variables ν_x and μ.

An interpretation of equation (42) can be given : the 2-D FT of the section is obtained by projection of the 3-D FT of the object on a plane section $(\nu_x 0\mu)$ in the Fourier domain. This projection is parallel to the plane $(\nu_y 0\nu_z)$ and weighted by the adequate Fourier coefficients (IFT_{1,ν_y}). The projection plane $(\nu_x 0\mu)$ is the corresponding section in the Fourier domain of the section $(x0\sigma)$, and the central frequency plane has the same obliquity as the corresponding real plane (Fig. 7).

We leave the equation (42) in this general expression but it will be transformed later. However, some examples of its use can be given for two particular cases.

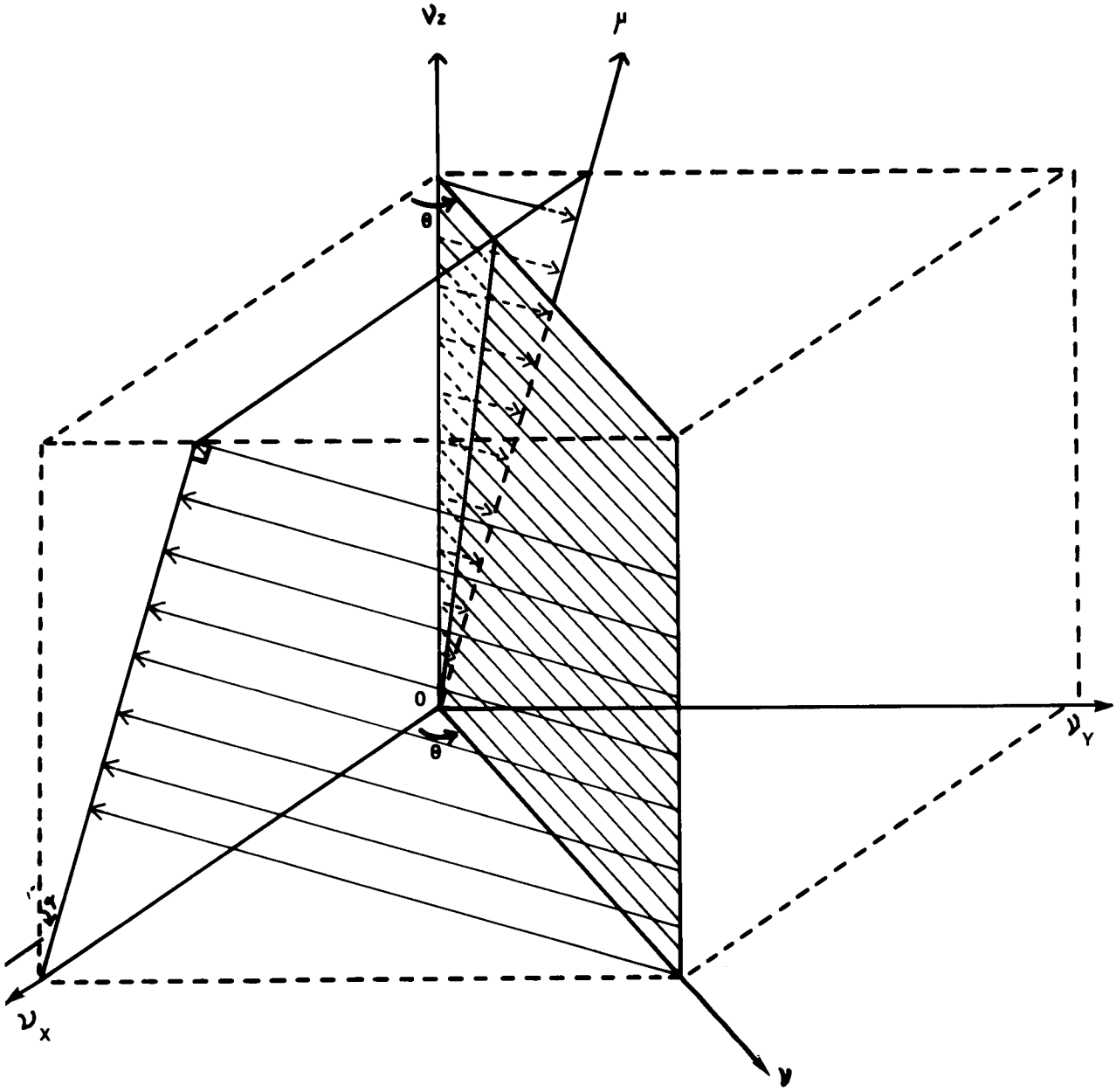


FIG. 7 - Projection of a central slice ($\nu 0 \nu_z$) in the Fourier domain on the Fourier oblique plane ($\nu_x 0 \mu$). The projection represented by arrows is parallel to the plane ($\nu_y 0 \nu_z$) and perpendicular to the Fourier oblique plane.

Particular Cases

To calculate (42), it is not necessary to know A explicitly. For example, if the ensemble of all the central slices containing the axis $(O\nu_z)$ are known, thus A is also known. Consider such a central slice A_θ ; its 3-D extension (see Appendix I) is defined in cylindrical coordinates by

$$\tilde{A}(\nu,\nu_z) = A(\nu,\varphi,\nu_z).\delta_\theta(\varphi) \text{ with } \nu = \sqrt{\nu_x^2+\nu_y^2} \tag{43}$$

and, by superimposition of all the extensions (see Appendix I), we get

$$SO\tilde{A}_\theta = \frac{1}{|\nu|}.A \tag{44}$$

Then

$$\begin{aligned} IFT_{1,\nu_y}(A)(\nu_x,\mu;d) &= IFT_{1,\nu_y}(|\nu|.SO\tilde{A}_\theta)(\nu_x,\mu;d) \\ &= SOIFT_{1,\nu_y}(|\nu|.\tilde{A}_\theta)(\nu_x,\mu;d) \end{aligned} \tag{45}$$

and a straightforward calculation gives

$$IFT_{1,\nu_y}(|\nu|.\tilde{A}_\theta)(\nu_x,\mu;d) = B_{\theta,\alpha}(\nu_x,\mu).\exp(2i\pi.\nu_x.d.tg\ \theta) \tag{46}$$

with

$$B_{\theta,\alpha}(\nu_x,\mu) = \left|\frac{\nu x}{\cos\theta}\right|.A_\theta\left[\frac{\nu x}{\cos\theta},\ \frac{\mu-\nu_x.tg\ \theta.\sin\alpha}{\cos\alpha}\right] \tag{47}$$

where $B_{\theta,\alpha}$ is the projection of $|\nu|.\tilde{A}_\theta(\nu,\nu_z)$ on the plane $(\nu_x O\mu)$. Therefore, $IFT_{1,\nu_y}(|\nu|.A_\theta)$ is the projection of $|\nu|.\tilde{A}_\theta(\nu,\nu_z)$ weighted by $\exp(2i\pi.\nu_x.d.tg\ \theta)$, on the plane $(\nu_x O\mu)$.

Another particular case occurs when d and α are equal to zero. Then (42) yields

$$f_d(X,Z) = IFT_2\left(\int_{-\infty}^{+\infty} A(\nu_x,\nu_y,\nu_z).d\nu_y\right)(X,Z) \tag{48}$$

This equation shows that each central slice of f containing the rotation axis can be obtained by the 2-D IFT of the projection of A on the central plane containing the corresponding axis $(O\nu_z)$ with the same rotation angle in the Fourier domain. We recognize there the reciprocal theorem of the Central Slice Theorem. The 3-D reconstruction equations with parallel projections are nothing else but particular cases of the equation (42), due to the preferenced rotation axis.

DETECTOR AND NOISE

An important distortion in the 3-D reconstruction is introduced by the spatial resolution of the scintillation camera. In fact, the different adjacent transverse sections are not independent in spite of the theory. The theoretical 2-D reconstruction algorithms must be modified and extended to three dimensions. The reconstruction of one section must use all the informations contained in the field of the scintillation camera. The problem of noisy data also introduces the same constraint. Therefore, it is necessary to perform a resolution correction and a noise filtering. The entire system of detection is assumed to be shift-invariant, that is completely characterized by a unique point-spread function

(PSF). This hypothesis gives good results in practice though wrong for a scintillation camera.

This PSF is different from the PSF of the motionless collimator and can be estimated from the 3-D reconstruction of a revolving point source. Thus, one also takes into account the loss of resolution due to the algorithms used. This PSF is then characteristic of the rotating system and of the analytical techniques of reconstruction. In the general case, this function is approximately an ellipsoïd.

The use of a filter h allows to perform simultaneously the resolution correction and noise filtering. Consider now the modifications introduced by this filter in the fundamental equations.

Transverse Sections

Consider A the 3-D FT of an object f. A transverse section can be defined by

$$z = Z_o \text{ for all x and y} \tag{49}$$

then, as for the oblique sections, the 3-D IFT of A passing through (X,Y,Z_o) is

$$\begin{aligned} f(X,Y,Z_o) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} A(\nu_x,\nu_y,\nu_z).\exp(2i\pi.\nu_z.Z_o)d\nu_z\right].\exp(2i\pi(\nu_x.X+\nu_y.Y)).d\nu_x.d\nu_y \\ &= IFT_2(IFT_{1,\nu_z}(A)(\nu_x,\nu_y;Z_o)) \end{aligned} \tag{50}$$

Be H the 3-D FT of a filter h. As h is assumed to be circularly symmetric around the rotation axis (Oz), it is the same for H around the axis $(O\nu_z)$.

$$H(\nu_x,\nu_y,\nu_z) = H(\nu,\nu_z) \text{ with } \nu = \sqrt{\nu_x^2+\nu_y^2} \tag{51}$$

In this case, the result of the reconstruction is not f but f convolved with h. Replacing A by the product of H and the 3-D FT of f in equation (50) yields

$$A(\nu,\varphi,\nu_z) = H(\nu,\nu_z).FT_3(f)(\nu,\varphi,\nu_z) \tag{52}$$

Thus, the central slice A_θ can be written

$$A_\theta(\nu,\nu_z) = H(\nu,\nu_z).FT_3(f)(\nu,\theta,\nu_z) \tag{53}$$

and, according to the Central Slice Theorem, it follows that

$$A_\theta(\nu,\nu_z) = H(\nu,\nu_z).FT_2(g_\theta)(\nu,\nu_z) \tag{54}$$

Then, using (44) and (54) in (50) we get

$$\begin{aligned} IFT_{1,\nu_z}(|\nu|.\tilde{A}_\theta)(\nu,\theta;Z_o) &= |\nu|.\int_{-\infty}^{+\infty} A_\theta(\nu,\nu_z).\exp(2i\pi.\nu_z.Z_o).d\nu_z \\ &= |\nu|.\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g_\theta(r,z).\exp(-2i\pi.(\nu.r+\nu_z.z)).dr.dz\right).H(\nu,\nu_z).\exp(2i\pi.\nu_z.Z_o).d\nu_z \\ &= |\nu|.\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} H(\nu,\nu_z).\exp(2i\pi.\nu_z.(Z_o-z)).d\nu_z\right).g_\theta(r,z).\exp(-2i\pi.\nu.r).dr.dz \\ &= |\nu|.\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} IFT_{1,\nu_z}(H)(\nu;Z_o-z).g_\theta(r,z).\exp(-2i\pi.\nu.r).dr.dz \end{aligned} \tag{55}$$

Assuming that H can be separated in two 1-D functions H_1 and H_2, so that

$$H(\nu,\nu_z) = H_1(\nu).H_2(\nu_z) \tag{56}$$

equation (55) can be written

$$\begin{aligned} IFT_{1,\nu_z}(|\nu|.\tilde{A}_\theta)(\nu,\theta;Z_o) &= |\nu|.H_1(\nu).\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} IFT_{1,\nu_z}(H_2)(Z_o-z).g_\theta(r,z).\exp(-2i\pi.\nu.r).dr.dz \\ &= |\nu|.H_1(\nu).\int_{-\infty}^{+\infty}(h_2 \underset{z}{*} g_\theta)(r;Z_o).\exp(-2i\pi.\nu.r).dr \end{aligned} \tag{57}$$

where $\underset{z}{*}$ denotes the 1-D convolution with respect to z and h_2 is the 1-D IFT of H_2. Equation (57) gives also

$$IFT_{1,\nu_z}(|\nu|.\tilde{A}_\theta)(\nu,\theta;Z_o) = FT_{1,r}(h_1' \underset{r}{*}(h_2 \underset{z}{*} g_\theta))(\nu,Z_o) \tag{58}$$

where $h_1' = IFT_1(|\nu|.H_1)$ and where $\underset{r}{*}$ denotes the 1-D convolution with respect to r. Equation (58) shows that, first $g_\theta \underset{z}{*} h_2$ is calculated at the points $(r;Z_o)$ (r variable) and afterwards the 1-D result is convolved with h_1'. The final line obtained, taken as a 1-D projection, is then treated by the usual 2-D techniques.

These results can be considered as a starting point to implement other techniques in the Fourier or in the real domains.

Oblique Sections

The filter h should be introduced in the general expressions written for the plane oblique sections. Equations (51) to (54) obtained for transverse sections can be used without modification. The $B_{\theta,\alpha}$ defined in (46) and (47) can be interpreted as the weighted projection of the 2-D FT of the 2-D convolution of g_θ and h' (with $FT_2(h') = |\nu|.H$).

The equivalence of the back-projection in the real domain and the projection weighted by $1/|\nu|$ in the Fourier domain, allows to smear $g_\theta * h'$ with the shift defined by $\delta_d.\sin\theta$. Therefore, to obtain the section of interest, we superpose the values at the intersections between the $(xO'\sigma)$ plane and the back-projection of each projection.

From the fundamental equations of the oblique sections and from particular cases, different algorithms have been deduced in the Fourier and in the real domains without reconstruction of the 3-D FT of the object.

For example, the equations can be simplified in using 2-D techniques. In this case, each oblique plane can be considered as an ensemble of parallel lines at different levels Z_o. Thus, the problem is limited to the reconstruction of a any line of a 2-D object. The algorithms defined by equation (50) or subsequent equations as (58), were first used to reconstruct a single line contained in the oblique plane of each transverse section. This technique assumes the separability of the variables in the frequency filter.

A faster but less accurate technique which neglects the resolution of the system, requires the computation of each line of interest of each transverse section. Then, a 1-D filter is applied to each line and, after the construction of the oblique plane, another filter is used along the (Oz) axis. This technique, in spite of its approximations, gives good results.

These two techniques reconstruct the plane oblique section projected along the (Oz) axis, that is with an angle of view $(\pi/2-\alpha)$ whereas the sections reconstructed by the general techniques are seen with a maximum angle of view $(\pi/2+\alpha)$. Thus, the simplified techniques are only efficient to detect defects

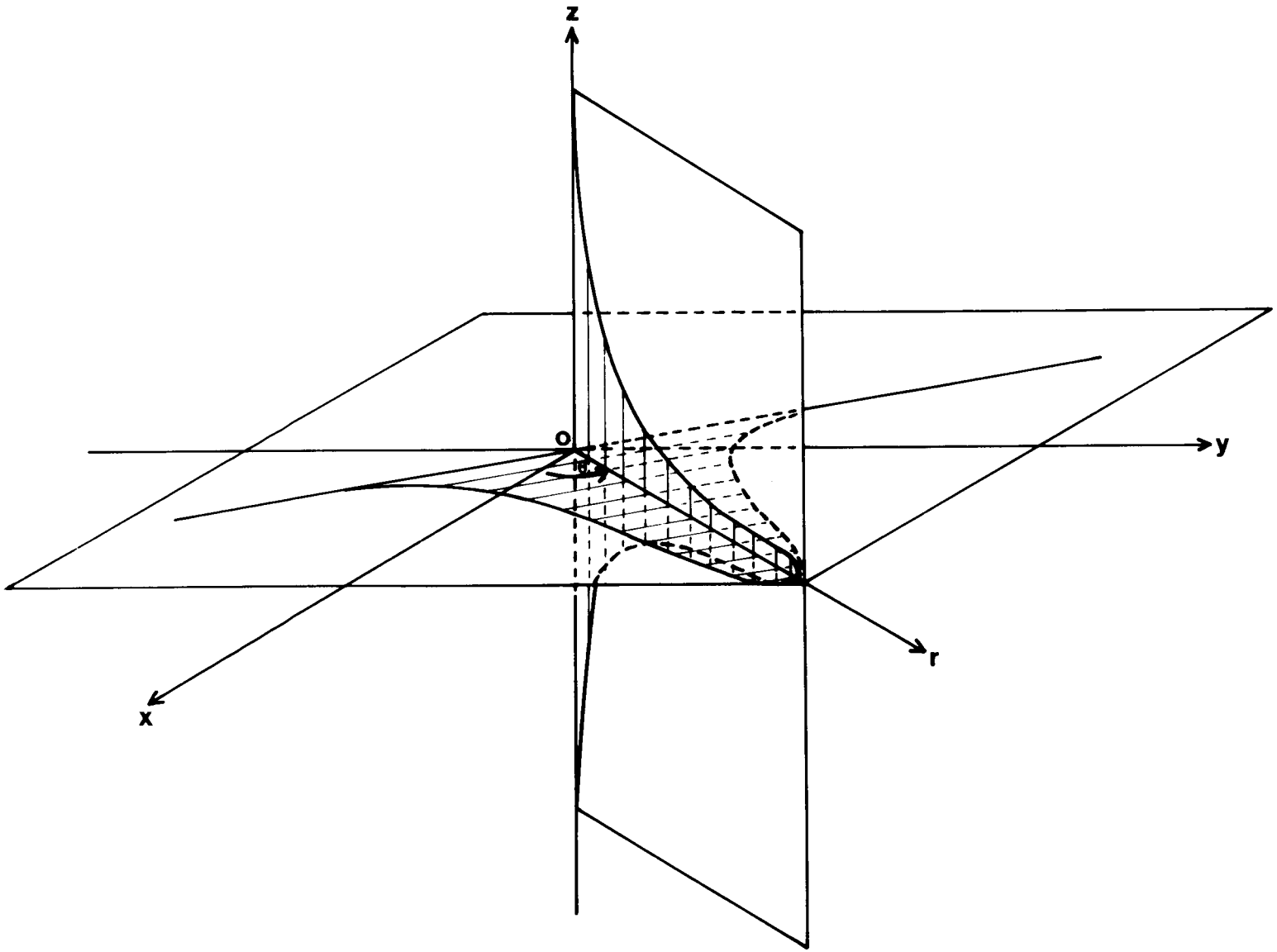


FIG. 8 - 3-D real response function represented by two Gaussian functions in the (xOy) and (rOz) planes.

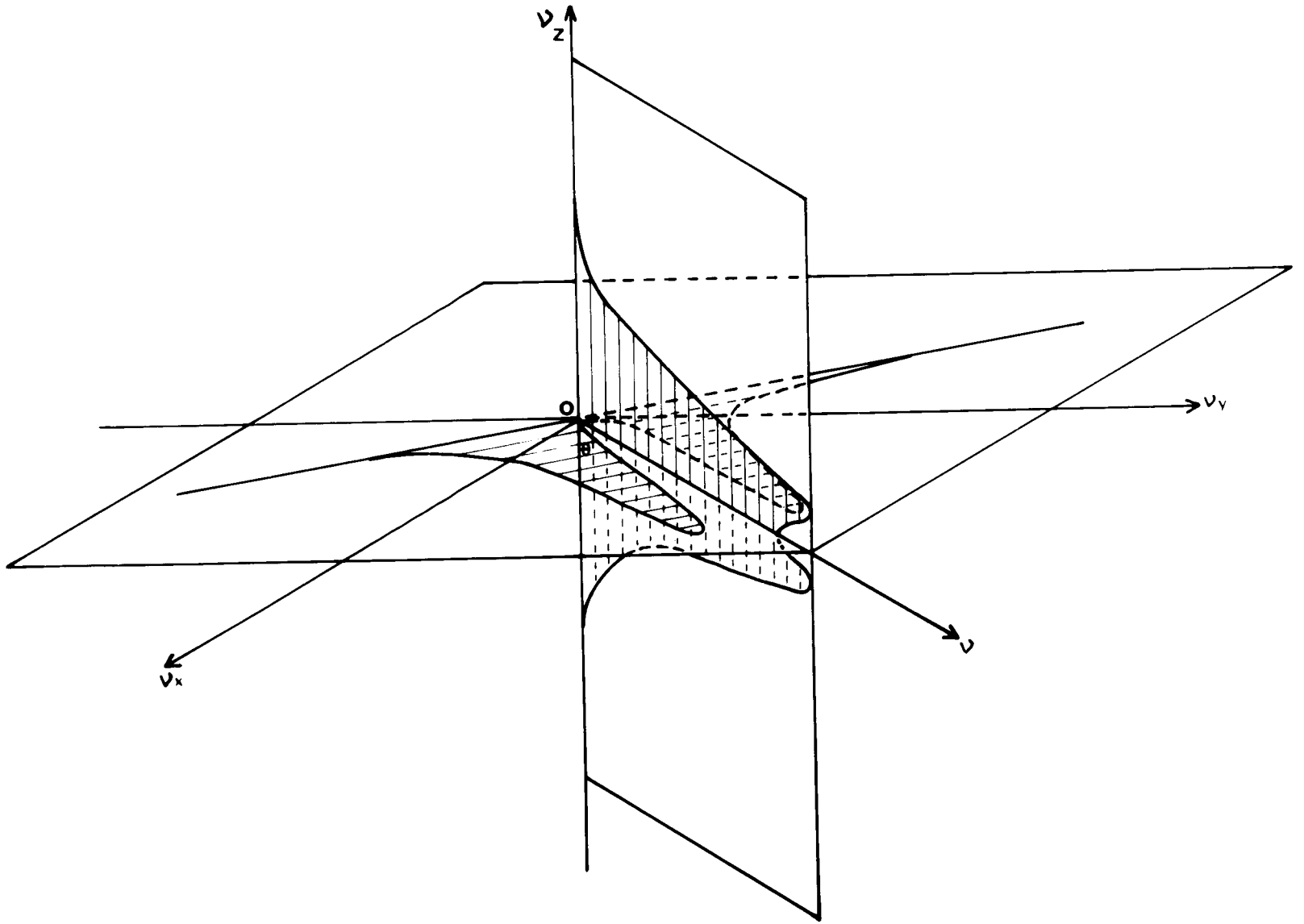


FIG. 9 - 3-D Wiener filter in the Fourier domain obtained with the two Gaussian functions of the figure 8.

when their projections on the (Oz) axis, are well separated. Their major interest, compared to the general techniques, is to divide by 5 to 10 the computation time. They are useful when α is quite different from $\pi/2$.

In practice, the resolution correction and the noise filtering are done using a Wiener filter[27] approximating the PSF by a product of two gaussian functions, the first along the rotation axis, the second in the rotation plane (Fig. 8). The figure 9 shows the corresponding filter used for the reconstruction.

RESULTS

Radioactive Phantom

To study the phantoms, we use a small rotating plate (1 turn/minute) placed in front of the scintillation camera ; its rotation axis being located at a distance of 20 cm to 30 cm from the collimator. To determine the true position of the rotation axis, it is necessary to collect the data from a rotating point source. To position the sections accurately one or more standard scintigrams are acquired.

The object was a Perspex box filled with 2,300 cm^3 of a solution containing 1 mCi of ^{99m}Tc. Three rectangular parallelepipeds (P1, P2, P3) filled with water were placed inside this box. The dimensions of their square base are : 4 cm (P1), 3.2 cm (P2) and 2.5 cm (P3) (Fig. 10). The rotation axis is placed 5 mm behind and 2 mm to the left of P1.

Two conventional scintigrams (200,000 counts) were performed with the Searle gamma camera (Fig. 11) in the initial position (A) and after a rotation of 180 degrees (B). From the list mode acquisition performed with the same camera, 33,000 counts (120 seconds) were used to reconstruct the transverse sections, of thickness 2.5 cm, of the picture 12-A and all the other sections of the figures 12 and 13 are reconstructed from the same 58,000 counts (60 seconds).

We think that the results obtained indicate that the rotation axis was not exactly parallel to the collimator face but that this fact did not introduced a major distortion.

The computation time for all these pictures varies from 2 to 10 minutes with the Univac 1106 for a 128x 128 matrix and a pixel size of 1.9 mm X 1.9 mm.

Clinical Studies

For the clinical studies, we modified an old rotating chair used in tomoradiology, to get a rotation speed of one turn per second to two turns per minute. We built a removable arm support to avoid tiring the patient too much. We have mainly studied livers because this kind of organ leads to the major problems in 3-D reconstruction due to the internal absorption and the respiratory motions. For the liver examination, the patient is sitting approximately at 50 cm in front of the collimator face to prevent his knees from touching the camera. This produces a significant loss of resolution (cf. Fig. 14, 16 and 18). The center of rotation is determined by means of a rotating point source and one or more classical images are recorded in order to position the sections. The tomographic record is performed by the Toshiba camera and 500,000 to 2,000,000 counts are collected in a maximum time of 10 minutes.

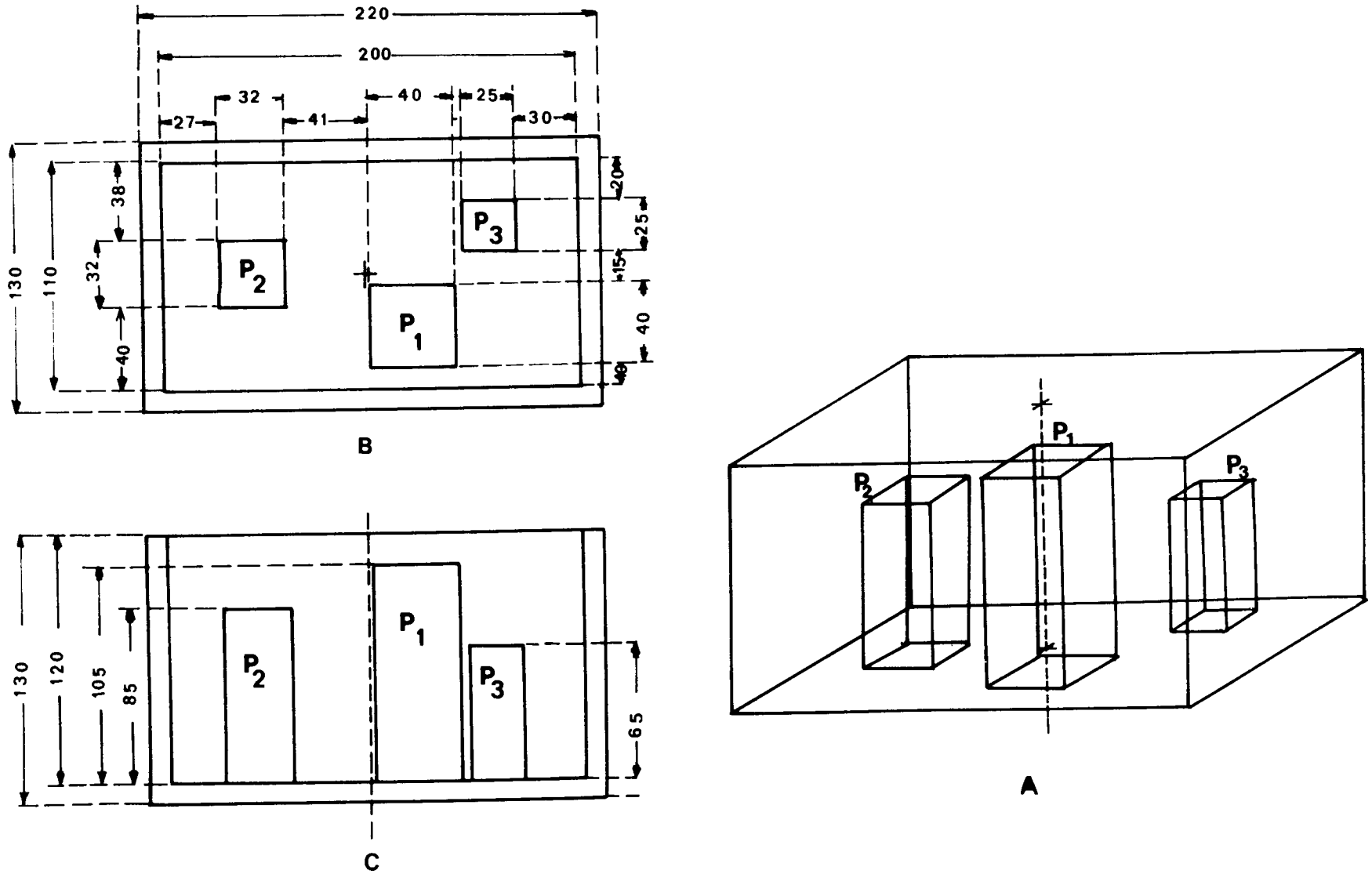
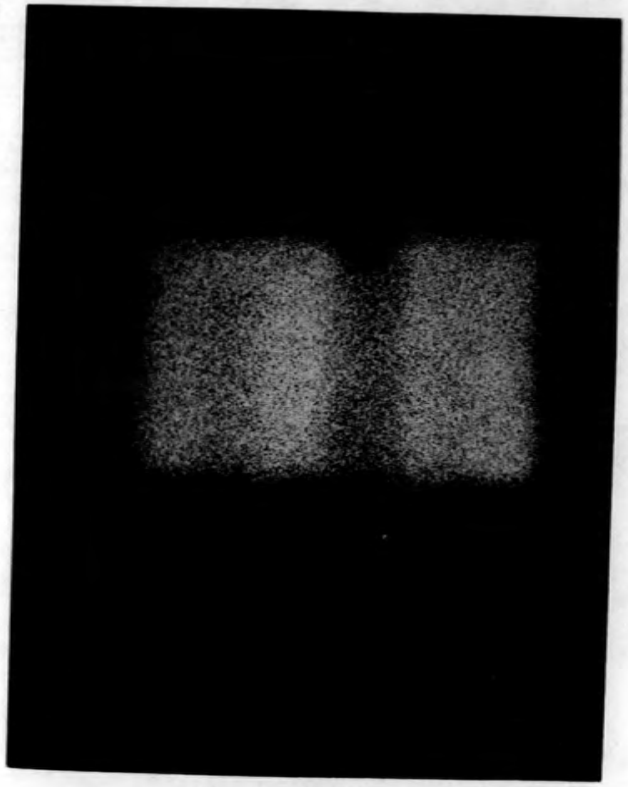
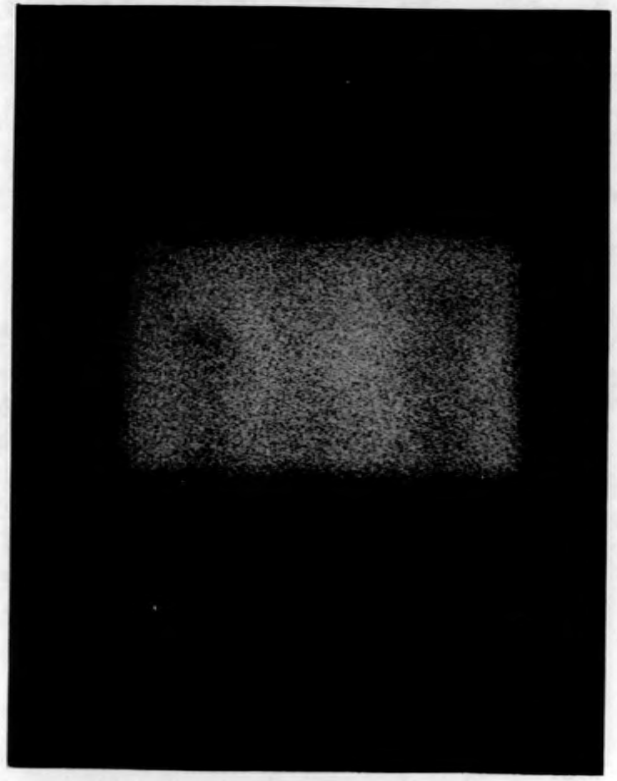


FIG. 10 - Diagram of the radioactive phantom used : 3-D view of the phantom (A), superior view (B), and anterior view (C). The rotation axis is represented by a cross (B) or a dashed line (A,C). Distances are given in millimeters.



A



B

FIG. 11 - Standard scintigrams of the phantom of the figure 10. Image obtained in the initial position of the 3-D record or anterior view (A). Image obtained after a rotation of 180 degrees or posterior view (B).

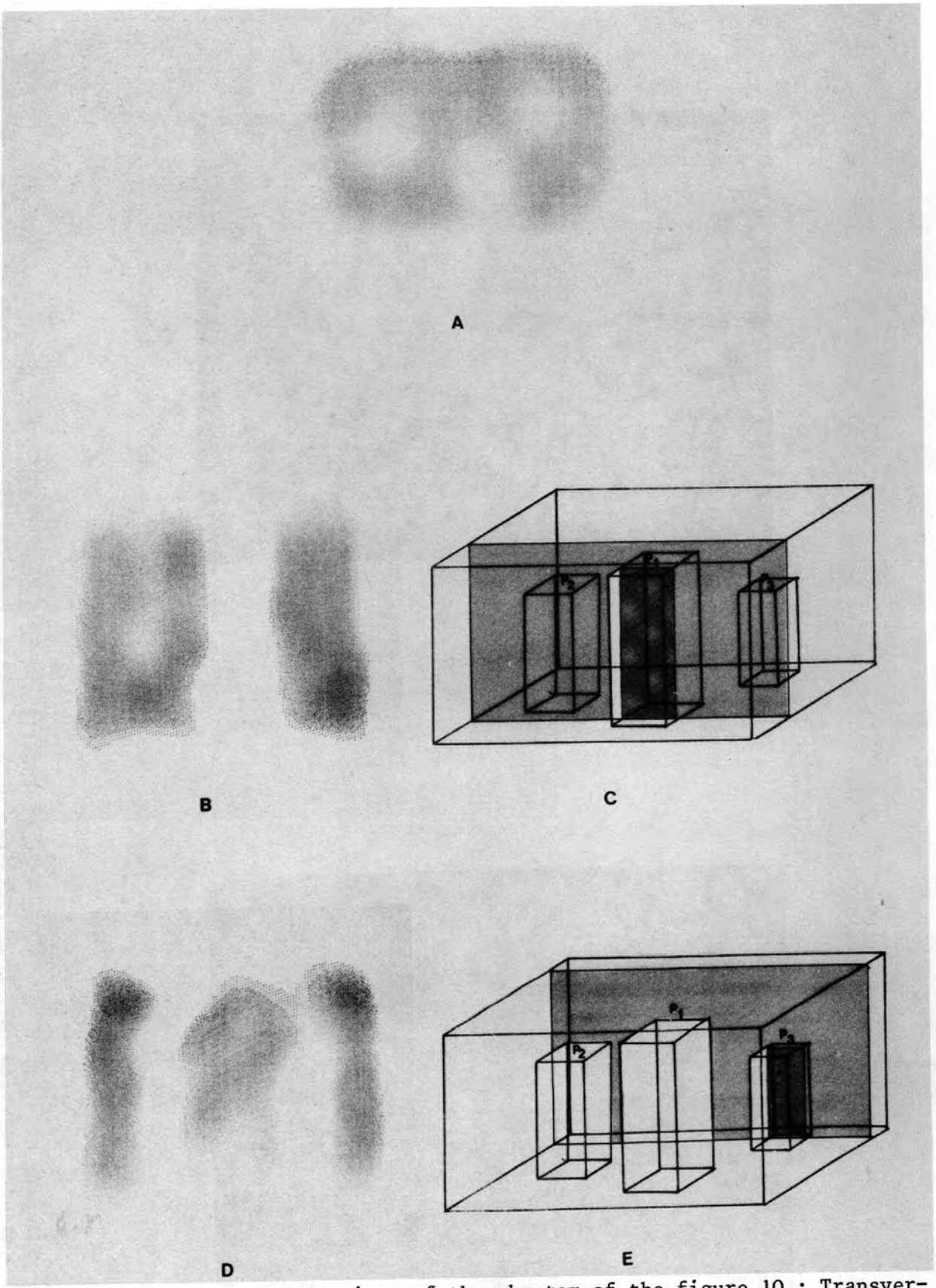


FIG. 12 - Reconstructed sections of the phantom of the figure 10 : Transverse section (A). Frontal section at 3 cm before the rotation axis (anterior view)(B). Frontal section at 3 cm behind the rotation axis, 1.7 cm behind P2, 3.5 cm behind P1 and through the posterior part of P3 (anterior view) (D). Pictures C and E are schematic drawings of sections B and D.

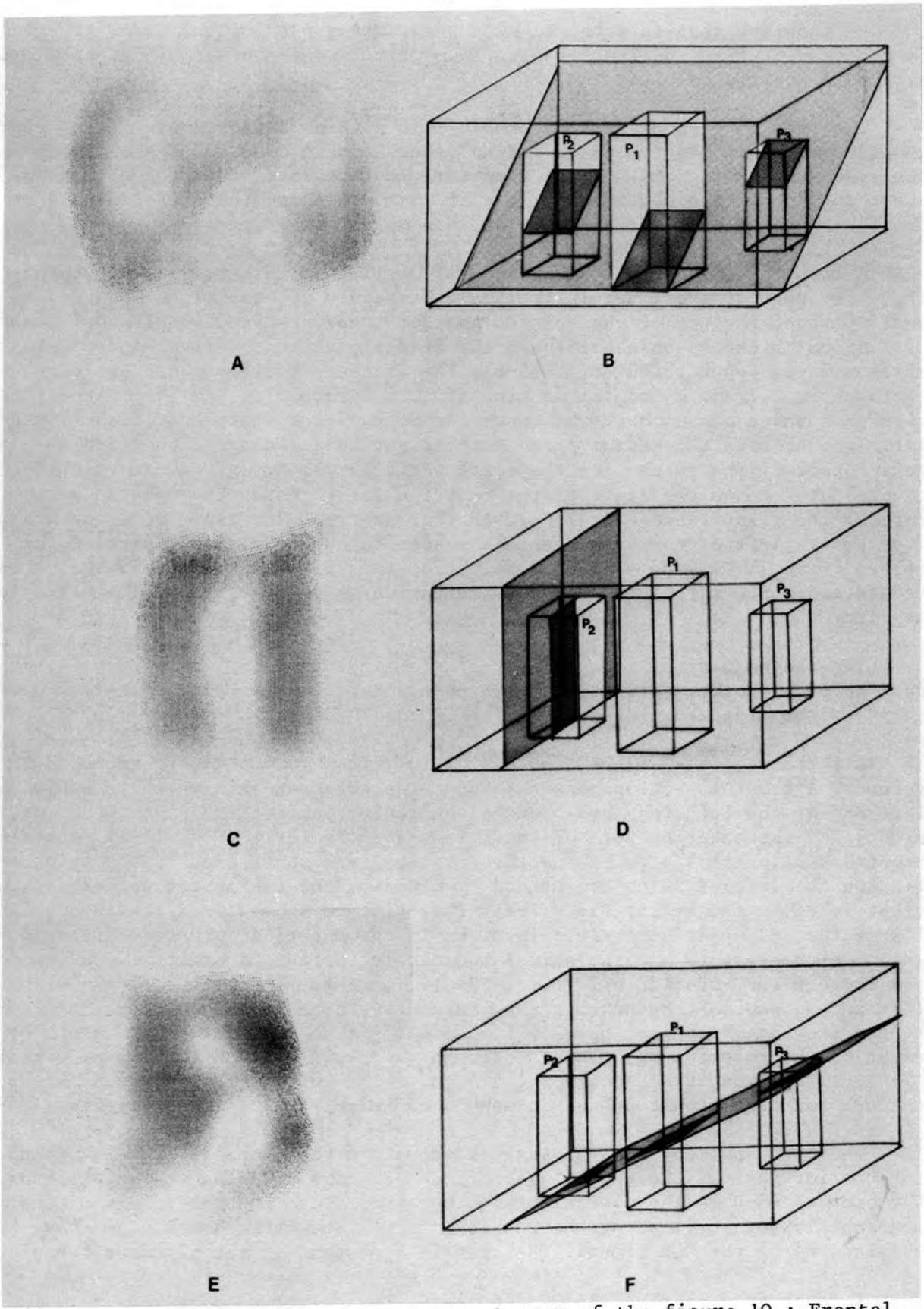


FIG. 13 - Reconstructed sections of the phantom of the figure 10 : Frontal oblique section where P3 seems to be connected to P1 (anterior view) (A). Sagittal section at 5 cm from the rotation axis passing through P2 (lateral view) (C). Sagittal oblique section (from the bottom to the top) passing through the lower of P2, the middle of P1 and the upper part of P3 (lateral view) (E). Pictures B, D, F are schematic drawings of sections A, C, E.

We show the results obtained with a rotation speed of two turns per minute in the case of three patients having respectively : a sub-normal liver, a cirrhosis and a hepatic cyst.

The first patient is a 23-year-old male with an osteosarcoma of the right shoulder, a small hepatomegalia and no other hepatic disease. Figure 14 shows the standard scintigrams. On the anterior view obtained under the conditions of the tomographic examination, the loss of resolution and the different aspect of the liver due to the sitting position, are appearant (Fig. 14-C).

A transverse section (30,000 counts) of 1.35 cm thickness passes through the upper part of the liver (Fig. 15-A). We mainly see the right lobe and a small part of the top of the spleen. Another transverse section (50,000 counts) of 1.35 cm thickness passes through the middle part of the liver and crosses the spleen in its upper third (Fig. 15-B). The distance between these two transverse sections is 2.2 cm. A frontal section (80,000 counts) at 1 cm behind the rotation axis which passes near the posterior edge of the sharp end of the liver left lobe crosses the medium part of the right lobe (on the right) and the anterior part of the spleen (on the left) (Fig. 15-C). A sagittal section (90,000 counts) at 7 cm to the right of the rotation axis, passes through the medium part of the right lobe (Fig. 15-E). Another sagittal section (90,000 counts) at 7 cm to the left of the rotation axis passes through the medium part of the spleen (Fig. 15-F). The estimated height of the spleen is 15 cm and the estimated antero-posterior width is 7 cm. The patient seems to present a small splenomegalia.

The second patient is a 38-year-old female with a biliary cirrhosis secondary to angiocholitic infections with portal hypertension and ascitis. Figure 16 shows the standard images.

All the sections of the figure 17 result from 70,000 counts recorded in 30 seconds. A frontal section passes through the rotation axis which is behind the left end of the left lobe and crosses the medium part of the right lobe (on the right) and the anterior part of the spleen (on the left) (Fig. 17-A). Another frontal section at 5.4 cm behind the rotation axis shows clearly the splenomegalia (on the left) : 20 cm height and 7 cm width, and the posterior part of the right lobe (on the right)(Fig. 17-B). This picture shows a radioactive region in its medium and upper part. This is probably the anterior curvature of the dorsal spine which presents a pathological radioactive uptake. A sagittal section passes through the rotation axis (Fig. 17-C). The left lobe of the liver is on the left and we can see, on the right, a radioactive area which probably corresponds to the radioactive spinal uptake. A second sagittal section at 5.4 cm on the right of the rotation axis passes through the right lobe in its medium part (Fig. 17-D). A third sagittal section at 5.4 cm on the left of the rotation axis passes through the medium part of the spleen (24 cm height and 11 cm width)(Fig. 17-E)

The third patient is a 35-year-old female with a hepatic cyst in the postero-superior part of the right lobe. Figure 18 shows the standard scintigrams. The picture 18-D is the posterior view obtained under the conditions of the tomographic examination using the Toshiba camera. The other scintigrams have been obtained using the CGR camera. The cyst is appearant on the pictures B and C.

A transverse section (80,000 counts/4 minutes) of 1.35 cm thickness in the upper part of the liver gives a picture which is difficult to interpret because of the rather complicated structure of this liver (Fig. 19-A). After the surgical operation, we found that the cyst had a diameter of 8 to 10 cm and that two thirds of it was into the upper posterior part of the right lobe. This cyst seems to us smaller in the classical images than in this section. It is even not visible in the anterior view in spite of its large volume. Moreover, we see

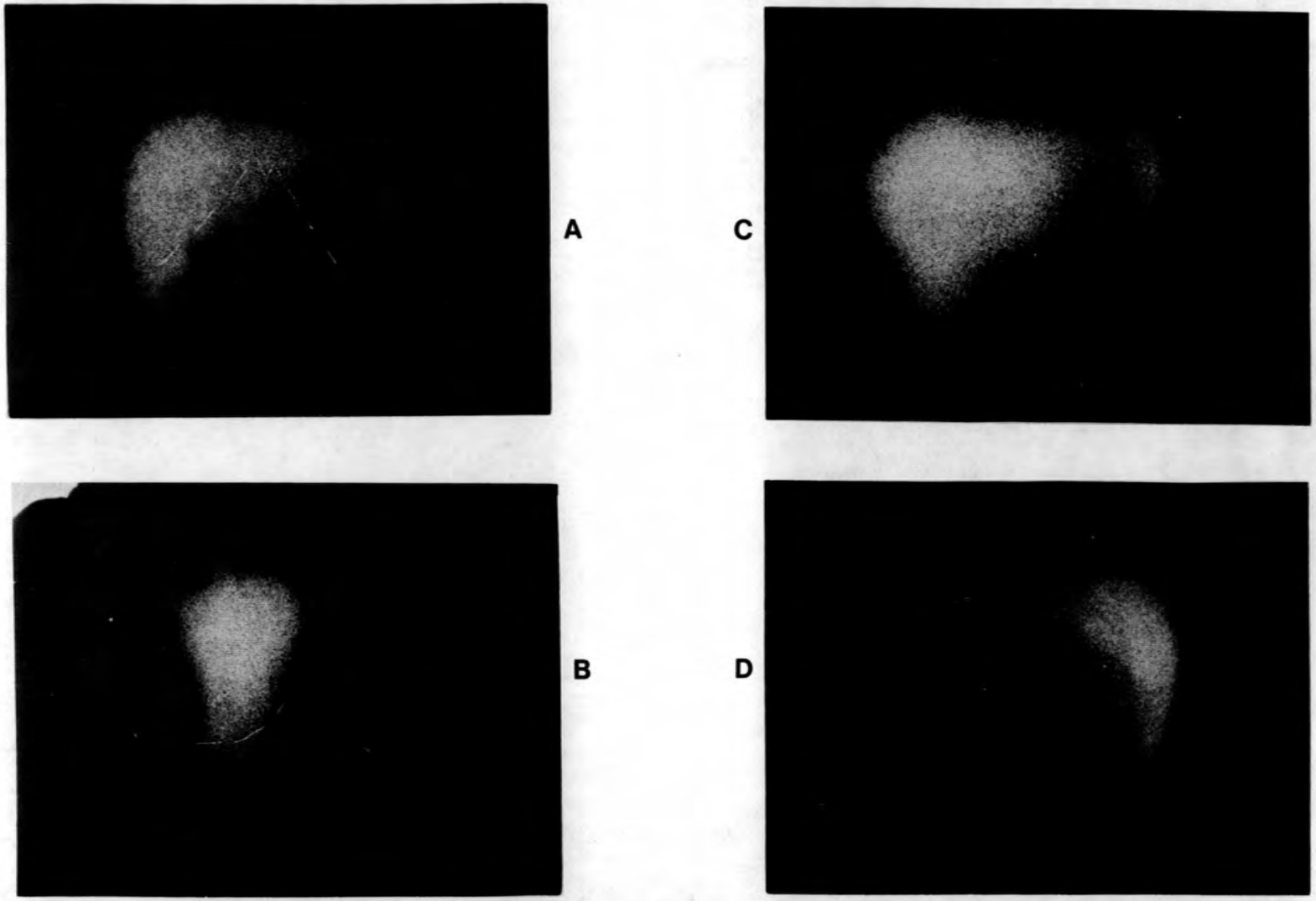


FIG. 14 - Standard scintigrams of a sub-normal liver of a patient with a small hepatomegalia : anterior (A), right lateral (B) and posterior (D) views. (C) is an anterior view obtained in the same conditions as the 3-D examination.

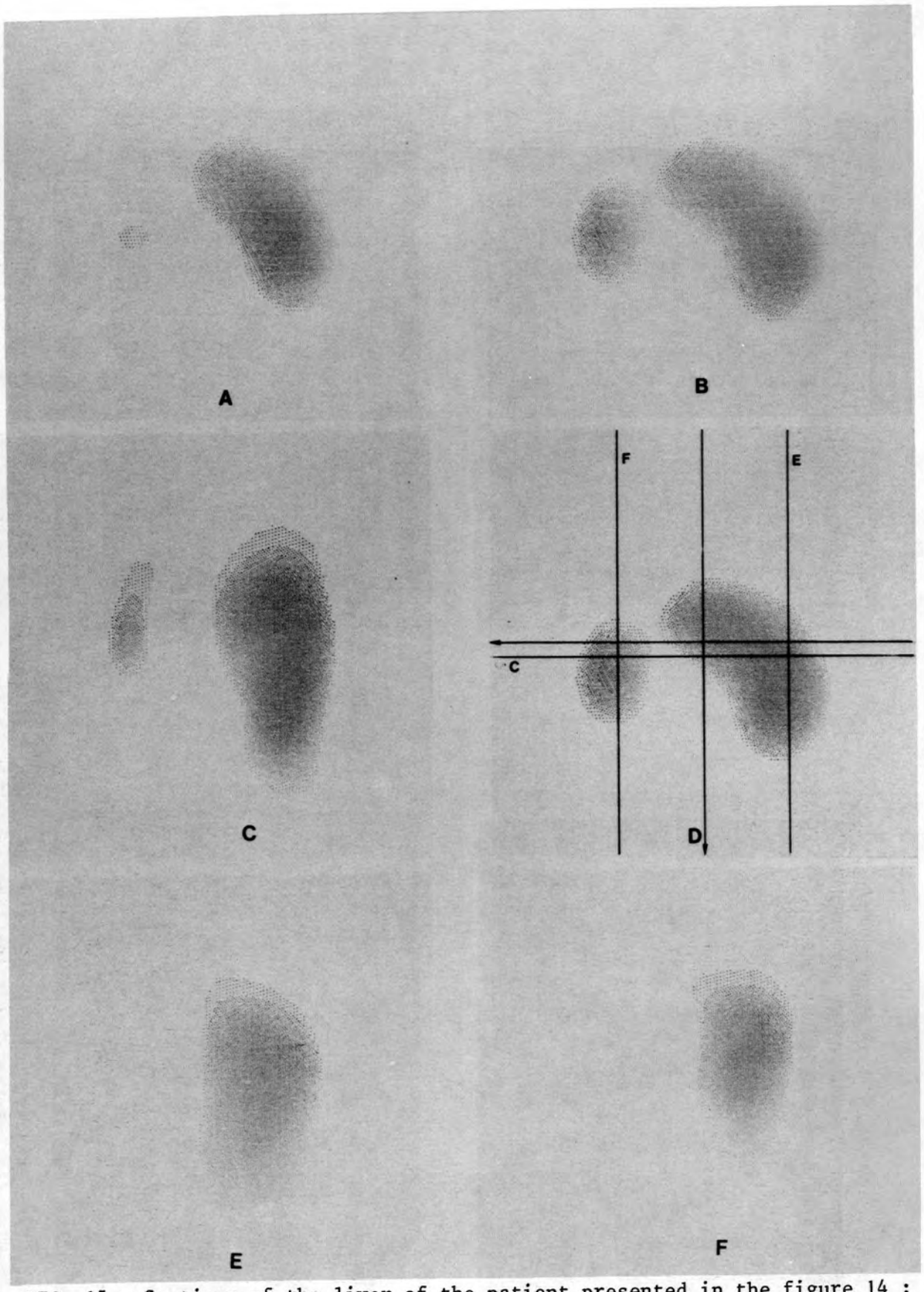


FIG. 15 - Sections of the liver of the patient presented in the figure 14 : transverse (A,B), frontal (C) and sagittal (E,F) sections. (A) and (B) are superior, (C) posterior, (E) and (F) left lateral (with the front on the left) views. The intersection lines of the sections C, E and F with the transverse section (B) are represented on the picture (D). The axes (Ox) and (Oy) are drawn with arrows, the initial axis (Ox) is horizontal.

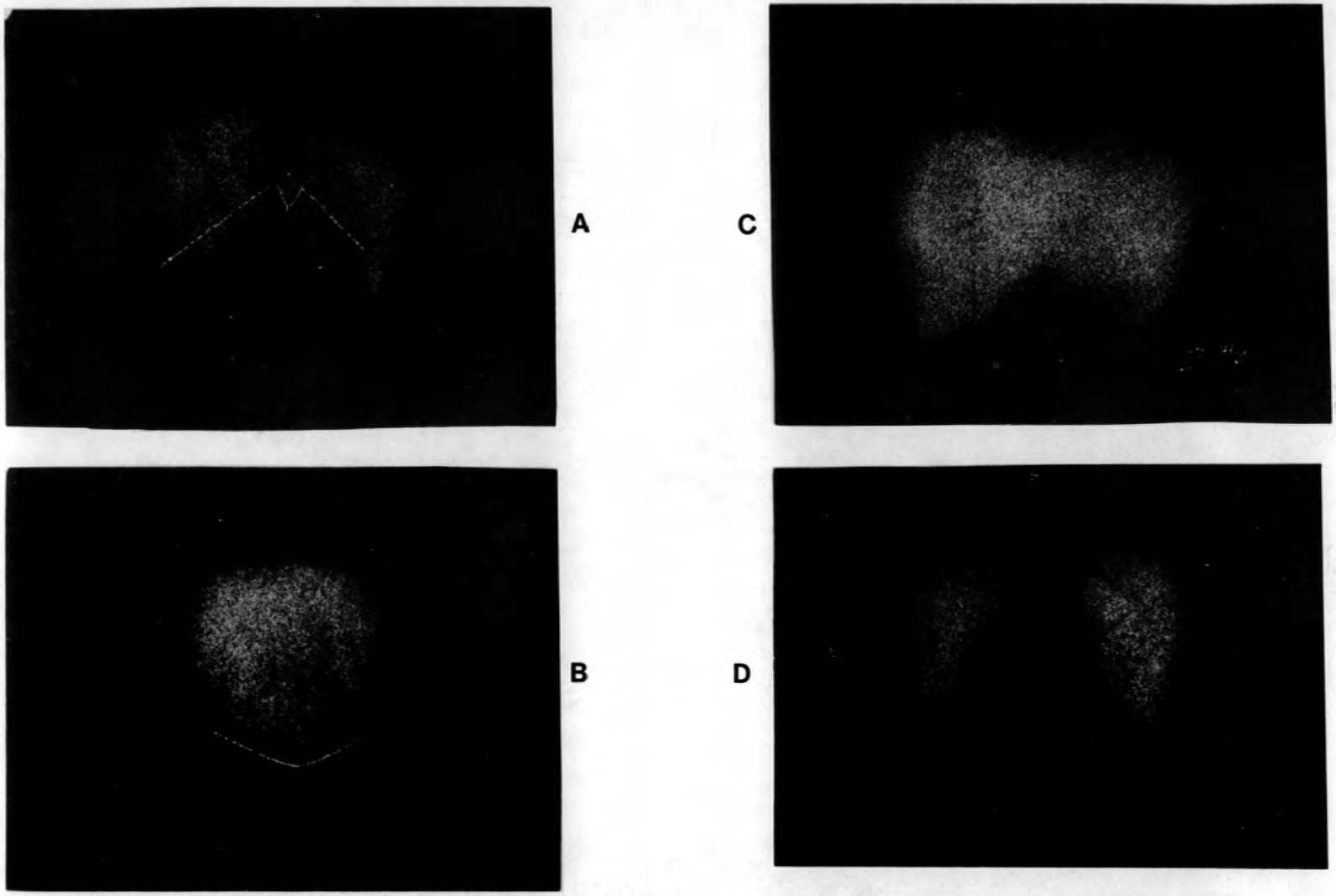


FIG. 16 - Liver standard scintigrams of a cirrhotic patient with hepatomegalia, splenomegalia, abnormal uptake of the spine and ascitis : anterior (A), right lateral (B) and posterior (D) views. The image (C) is an anterior view obtained in the same conditions as the 3-D examination.

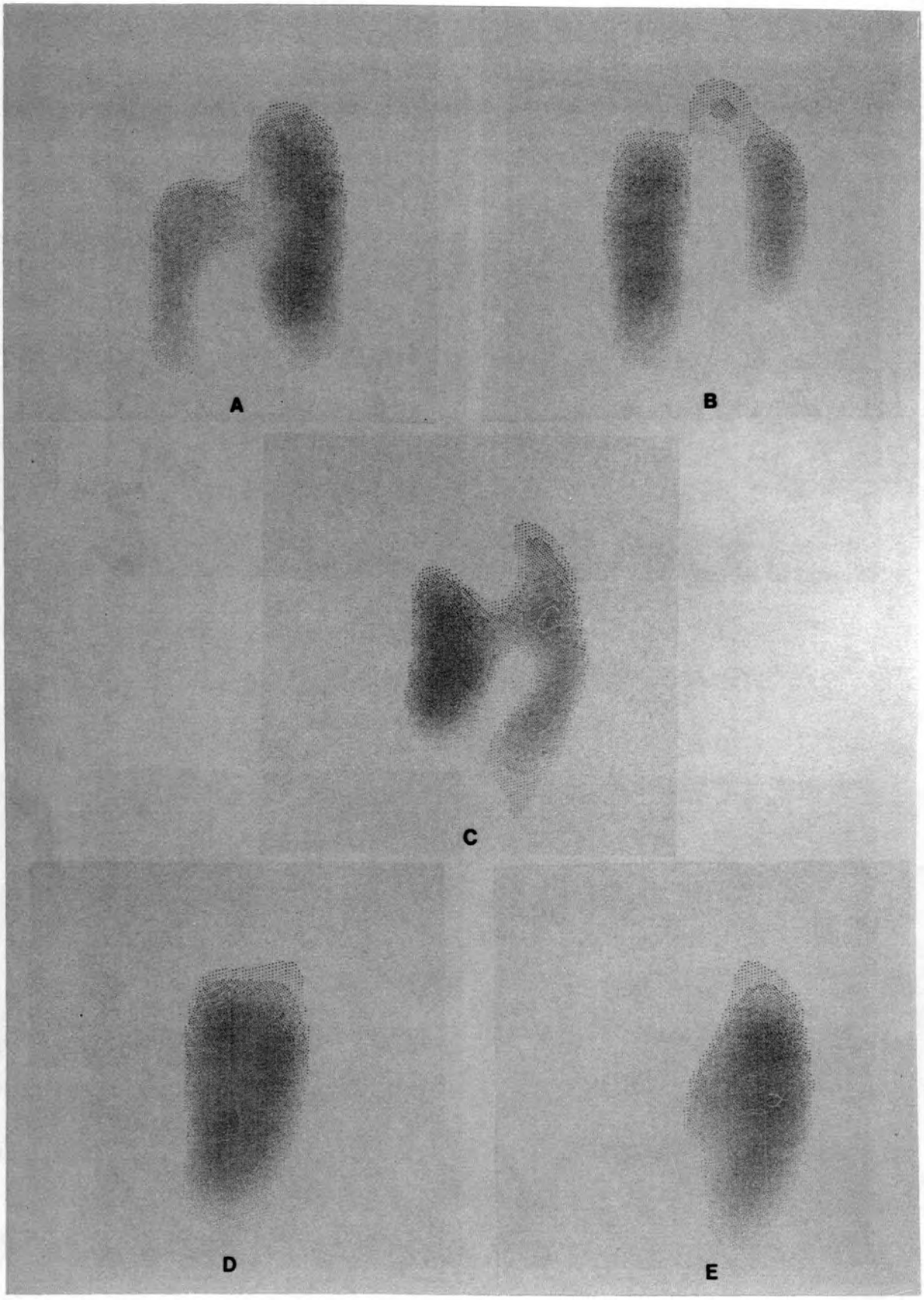


FIG. 17 - Sections of the liver of the patient presented in the figure 16 : frontal (A,B) (posterior view) and sagittal (C,D,E) (left lateral view) sections.

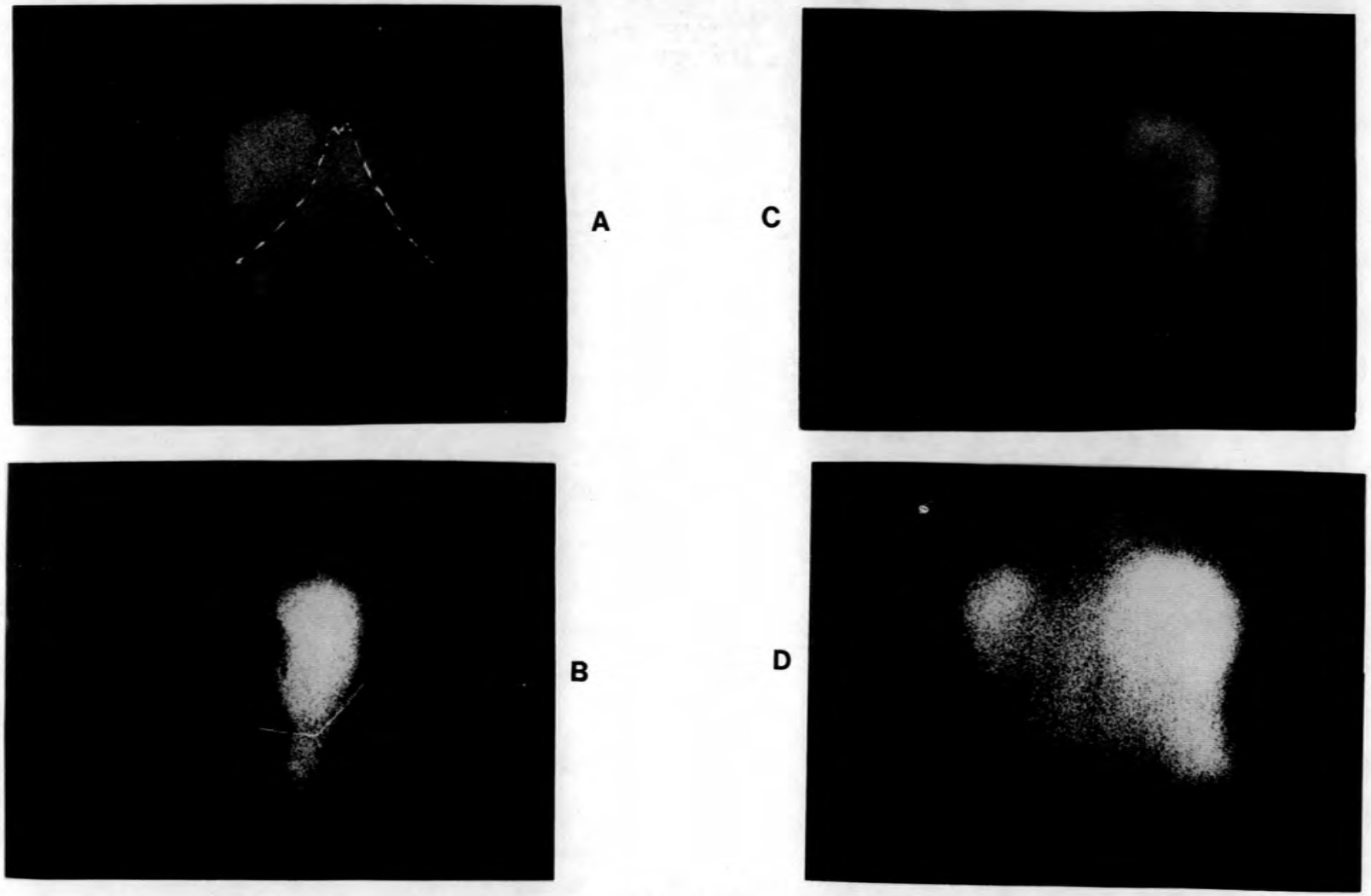


FIG. 18 - Liver standard scintigrams of a patient with a cyst in the postero-superior part of the right lobe : anterior (A), right lateral (B) and posterior (C) views. The image (D) is a posterior view obtained in the same conditions as the 3-D examination.

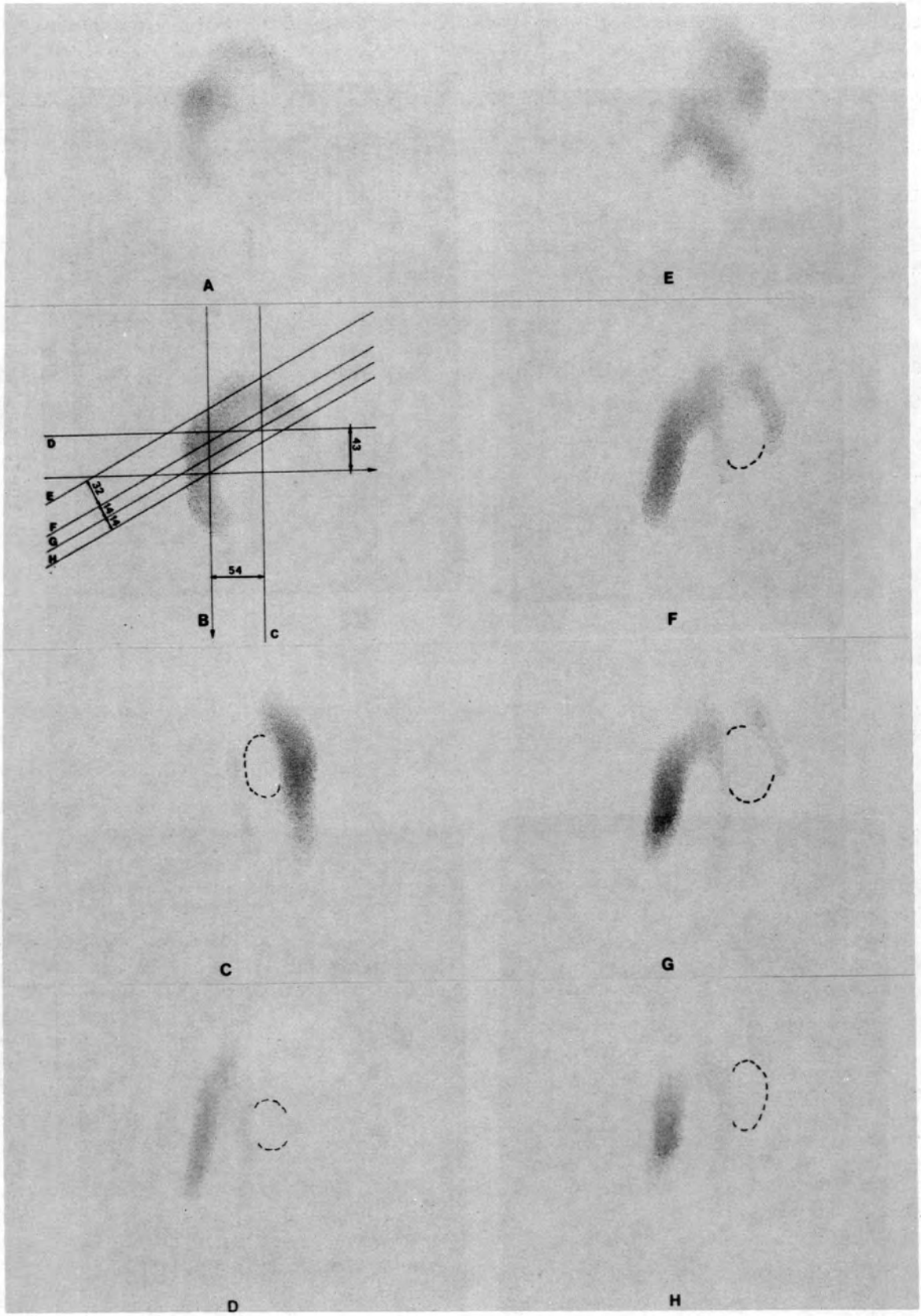


FIG. 19 - Sections of the cystic liver presented in the figure 18 : transverse (A) (superior view with the front to the left), frontal (C) (posterior view) and sagittal (D) (left-lateral view) through the cyst. The sections E, F, G, H are perpendicular to the corresponding lines drawn on the picture B, which is identical to A, and where the distances are given in millimeters and the initial axis (Ox) is vertical. The right lobe is on the right side. The dashed lines outline approximately the cyst.

the top of the spleen at the bottom. The sections C to H (170,000 counts/60 seconds) are all parallel to the rotation axis. A frontal section at 4.3 cm behind the rotation axis, passes through the cyst. The thin normal lateral part of the liver is on the right and the cyst is seen as a defect in the middle of the right lobe (Fig. 19-C). A sagittal section passes at 4.3 cm on the right of the rotation axis. The cyst is at the posterior part of the right lobe (on the right)(Fig. 19-D). On the two pictures C and D, the cyst cannot be seen clearly. To see it, different sections with a rotation angle of 60° with respect to the initial position were used : one section passes through the rotation axis and through the posterior part of the cyst (Fig. 19-H). Two other sections pass also through the cyst at a more inferior part (Fig. 19-F, G) and the last section passes through the anterior normal part of the liver (Fig. 19-E). Here we clearly see the cyst as a defect in the right lobe. For all the pictures of this figure, the cut-off frequencies of the Wiener filters used are twice as much as the cut-off frequencies of the filters used for the sections obtained from the other patients.

The computation time of each clinical section presented varies from 4 to 10 minutes on the Univac 1106 for a final matrix 128 x 128 and a pixel size of 2.7 mm x 2.7 mm.

CONCLUSIONS

Our studies with radioactive phantoms and patients demonstrate the clinical interest of emission tomography using a standard scintillation camera. However, various practical problems are still not completely solved :

1 - The device can be improved through the use of a special collimator with a good depth resolution and by construction of a system with continuous rotation of the camera.

2 - The computation time varies according to the techniques used between 2 and 60 minutes on the Univac 1106. In order to use such techniques routinely, an array processor, for example, would be necessary to divide the computation time by 10 to 100. Moreover, with a very large core memory it is possible to reconstruct the full 3-D object with a similar computation time.

3 - We presently study the problem of the internal absorption correction but it is obvious that only clinical studies will tell us if such a correction, which consists in adding the values of opposite projections, is of practical interest. Our results show that the problem of the internal absorption seems to be less crucial than it was foreseen.

4 - If the organ motions are very important, it is possible to record a physiological signal allowing to put the section of interest in the right position when it moves in the field of view of the camera. The transverse section considered must remain in the field of view during such motions.

REFERENCES

1. H.O. ANGER, D.C. PRICE and P.E. YOST, "Transverse-Section Tomography with the Scintillation Camera", J. Nucl. Med., 8, 314-315, 1967.

2. G. MUEHLLEHNER and R.A. WETZEL, "Section Imaging by Computer Calculation", J. Nucl. Med., 12, 2, 76-84, 1971.

3. T.F. BUDINGER and G.T. GULLBERG, "Three-Dimensional Imaging from Multiple Gamma-Camera Views", Proc. 1st World Congress of Nuclear Medicine, Tokyo, 1267-1272, 1974.

4. D.B. KAY, J.W. KEYES, Jr., and W. SIMON, "Radionuclide Tomographic Image Reconstruction Using Fourier Transform Techniques", J. Nucl. Med., 15, 11, 981-986, 1974.

5. J.W. KEYES, Jr., N. ORLANDEA, et al., "The Humongotron - A Scintillation - Camera Transaxial Tomograph", J. Nucl. Med., 18, 4, 381-387, 1977.

6. R.J. JASZCZAK, P.H. MURPHY, et al., "Radionuclide Emission Computed Tomography of the Head with 99m Tc and a Scintillation Camera", J. Nucl. Med., 18, 4, 373-380, 1977.

7. G. MUEHLLEHNER, "A Tomographic Scintillation Camera", Phys. Med. Biol., 16, 1, 87-96, 1971.

8. G.S. FREEDMAN, "Gamma Camera Tomography : Theory and Preliminary Clinical Experience", Radiology, 102, 365-369, 1972.

9. M. MATSUO, Y. INOUE, et al., "Study on Tomographic Radioisotopic Imaging with a Scintillation Camera", in Proc. 1st World Congress of Nuclear Medicine, Tokyo, 1275-1277, 1974.

10. P. SCHMIDLIN, "Three-Dimensional Scintigraphy with an Anger Camera and a Digital Computer", in Medical Radioisotope Scintigraphy (1972), Vol. 1, I.A.E.A., Vienna, 409-417, 1973.

11. R. DI PAOLA, J.P. BAZIN and M. DI PAOLA, "Improvement of Gamma-Camera Data Processing due to Sequential Acquisition", in Third International Conference on Data Handling and Image Processing in Scintigraphy, Boston, June 1973, 115-124, 1975.

12. R. DI PAOLA, J.P. BAZIN, et al., "Système en Ligne Adapté au Traitement des Informations Scintigraphiques", in Medical Radioisotope Scintigraphy 1972, Vol. 1, I.A.E.A., Vienna, 571-587, 1973.

13. H.H. BARRETT, S.K. GORDON and R.S. HERSHEL, "Statistical Limitations in Transaxial Tomography", Comput. Biol. Med., 6, 307-323, 1976.

14. A.M. CORMACK, "Reconstruction of Densities from their Projections, with Applications in Radiological Physics", Phys. Med. Biol., 18, 2, 195-207, 1973

15. D.J. DE ROSIER, A. KLUG, "Reconstruction of Three-Dimensional Structures from Electron Micrographs", Nature, 217, 130-134, 1968.

16. G.N. RAMACHANDRAN and A.V. LAKSHMINARAYANAN, "Three-Dimensional Reconstruction from Radiographs and Electron Micrographs : II. Application of Convolutions instead of Fourier Transforms", Proc. Nat. Acad. Sci., 68, 9, 2236-2240, 1971.

17. Z.H. CHO, J.K. CHAN, et al., "A Comparative Study of 3-D Image Reconstruction Algorithms with Reference to Number of Projections and Noise Filtering", IEEE Trans. on Nucl. Sci., 22, 344-358, 1975.

18. R. GORDON and G.T. HERMAN, "Three-Dimensional Reconstruction from Projections : A review of Algorithms", Int. Rev. Cyt., 38, 111-151, 1974.

19. R.A. BROOKS and G. DI CHIRO, "Principles of Computer Assisted Tomography (CAT) in Radiographic and Radioisotopic Imaging", Phys. Med. Biol., 21, 5, 689-732, 1976.

20. R.M. MERSEREAU and A.V. OPPENHEIM, "Digital Reconstruction of Multidimensional Signals from their Projections", in Proceedings of the IEEE, October 1974, 1319-1338, 1974.

21. J. RADON, "Uber die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten", Ber. Saechs. Akad. Wiss. Leipzig, Math-Phys., 69, 262-277, 1917.

22. R.A. CROWTHER, D.J. DE ROSIER and A. KLUG, "The Reconstruction of a Three-Dimensional Structure from Projections and its Applications to Electron Microscopy", Proc. Roy. Soc. Lond., 317, 319-340, 1970.

23. D.G. Mc CAUCHEY and R.P. KRUGER, "Fourier-Bessel Methods for Transverse-Axial Reconstruction", in Image Processing Research, Semi-Annual Technical Report, 1 March 1974 to 31 August 1974, USCIPI Report 540, 1974.

24. S.C. PANG and S. GENNA, "A Fourier Convolution Fan Geometry Reconstruction Algorithms", in Image Processing for 2-D and 3-D Reconstruction from Projections, Stanford University Institute for Electronics in Medicine and Optical Society of America, August 4-7, 1975.

25. G.T. HERMAN, A.V. LAKSHMINARAYANAN and A. NAPARSTEK, "Convolution Reconstruction Techniques for Divergent Beams", Comput. Biol. Med., 6, 259-271, 1976.

26. B.E. OPPENHEIM, "Three-Dimensional Reconstruction from Incomplete Projections", in Information Processing on Scintigraphy, Proceedings of the Fourth International Conference, Orsay, France, July 1975, 288-324, 1975.

27. C. BERCHE and R. DI PAOLA, "Analyse Fréquentielle et Filtrage des Images Scintigraphiques. III-Filtres Linéaires Invariants de Correction de Résolution", Journal de Biologie et Médecine Nucléaire, Vol. IX, 38, 8-41, 1974

APPENDIX I

2-D Extension of a 1-D Function

A 2-D function can be defined, in polar coordinates, for example when a 1-D function g is placed along a line passing through the center of the reference axes with an angle θ_o. This is equivalent to surround g with zeros in the 2-D space, θ_o being the angle between g and the reference axis. This new 2-D function is denoted g_{θ_o} and called "2-D extension" (by abbreviation "extension") of g

$$\begin{aligned} \tilde{g}_{\theta_o}(r,\theta) &= g(r) \quad \text{if } \theta = \theta_o \\ \tilde{g}_{\theta_o}(r,\theta) &= 0 \quad \text{if } \theta \neq \theta_o \end{aligned} \tag{59}$$

This can be written

$$\tilde{g}_{\theta_o}(r,\theta) = g(r).\delta_{\theta_o}(\theta) \tag{60}$$

where δ_{θ_o} is the Dirac's distribution.

Superposition Principle

Let f represent a 2-D function. In cartesian coordinates, we can write

$$f(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x',y').\delta(x-x',y-y').dx'.dy' \tag{61}$$

In polar coordinates, we get

$$f(r,\theta) = \int_{0}^{\pi}\int_{-\infty}^{+\infty} f(r',\theta').\delta(r-r',\theta-\theta').r'.dr'.d\theta' \tag{62}$$

$$f(r,\theta) = |r|.\int_{0}^{\pi} f(r,\theta').\delta(\theta-\theta').d\theta' \tag{63}$$

This equation means that the density of $\delta(\theta-\theta')$ with respect to θ' is $1/|r|$. It is also the equation of the central slice of f for an angle θ, which may be written

$$f_\theta(r) = f(r,\theta) \text{for a fixed } \theta \text{ and r variable} \tag{64}$$

The extension, for an angle θ_o, of a central slice, for θ', is a 2-D function

$$(\tilde{f}_{\theta'})_{\theta_o}(r,\theta) = f(r,\theta').\delta_{\theta_o}(\theta) \tag{65}$$

If $\theta_o = \theta'$, we get the extension with a simplified notation

$$\tilde{f}_{\theta'}(r,\theta) = f(r,\theta').\delta_{\theta'}(\theta) \tag{66}$$

or

$$\begin{aligned} \tilde{f}_{\theta'}(r,\theta) &= f(r,\theta') \quad \text{if } \theta = \theta' \\ \tilde{f}_{\theta'}(r,\theta) &= 0 \quad \text{if } \theta \neq \theta' \end{aligned} \tag{67}$$

We call "superposition" of all these central slices, the addition, in the proper position, of the extensions of the ensemble of the central slices. Then, one obtains a new 2-D function called "superposed function" of f and denoted $\mathrm{So}\tilde{f}_\theta$

$$(\mathrm{So}\tilde{f}_\theta)(r,\theta) = \int_0^{\pi} \tilde{f}_{\theta'}(r,\theta).d\theta' = \int_0^{\pi} f(r,\theta').\delta(\theta-\theta').d\theta' \tag{68}$$

and, with (63)

$$(\mathrm{So}\tilde{f}_\theta)(r,\theta) = \frac{1}{|r|}.f(r,\theta) \tag{69}$$

In other words, the superposition of all the central slices of a 2-D function f gives a new 2-D function equal to the function f weighted by the inverse of the module of the distance to the center $1/|r|$. That is

$$\mathrm{So}\tilde{f}_\theta = \frac{1}{|r|}.f \tag{70}$$

2-D Repetition of a 1-D Function

Placing a 1-D function g into the 2-D space (extension of g for θ_o) and repeating this function parallely to itself (angle θ_o), we can generate another 2-D function. This "2-D repetition" (by abbreviation "repetition") of g is denoted $\hat{g}_\theta$ and defined by

$$\hat{g}_{\theta_o}(r,\theta) = \int_{-\infty}^{+\infty} g(r').\delta(r'-r.\cos(\theta-\theta_o)).dr' = g(r.\cos(\theta-\theta_o)) \tag{71}$$

The direct relation between the notions of repetition and extension will be emphasised in the next paragraph.

As for the extension, we can define the repetition, for an angle θ_o, of a central slice for θ' by

$$\widehat{(f_{\theta'})}_{\theta_o}(r,\theta) = \int_{-\infty}^{+\infty} f_{\theta'}(r').\delta(r'-r.\cos(\theta-\theta_o)).dr' = \int_{-\infty}^{+\infty} f(r',\theta').\delta(r'-r.\cos(\theta-\theta_o)).dr' \tag{72}$$

If $\theta_o = \theta'$, we get the repetition with a simplified notation

$$\hat{f}_{\theta'}(r,\theta) = \int_{-\infty}^{+\infty} f(r',\theta').\delta(r'-r.\cos(\theta-\theta')).dr' \tag{73}$$

2-D FT of a 1-D Function

Let g be a 1-D function and $\tilde{g}_{\theta_o}$ its 2-D extension for θ_o. Then, the 2-D FT of this extension is

$$\begin{aligned} FT_2(\tilde{g}_{\theta_o})(\nu,\varphi) &= \int_0^{\pi}\int_{-\infty}^{+\infty} g_{\theta_o}(r,\theta).\exp(-2i\pi.\nu.r.\cos(\theta-\varphi)).|r|.dr.d\theta \\ &= \int_0^{\pi}\int_{-\infty}^{+\infty} g(r).\delta_{\theta_o}(\theta).\exp(-2i\pi.\nu.r.\cos(\theta-\varphi)).|r|.dr.d\theta \\ &= \int_{-\infty}^{+\infty}\left(\int_0^{\pi}\delta_{\theta_o}(\theta).\exp(-2i\pi.\nu.r.\cos(\theta-\varphi)).d\theta\right).g(r).|r|.dr \end{aligned} \tag{74}$$

and, according to (63) where $\delta_{\theta_o}(\theta)$ has the density $1/|r|$ with respect to θ

$$\begin{aligned} FT_2(\tilde{g}_{\theta_o}(\nu,\varphi) &= \int_{-\infty}^{+\infty}\frac{1}{|r|}.\exp(-2i\pi.\nu.r.\cos(\theta_o-\varphi)).g(r).|r|.dr \\ &= \int_{-\infty}^{+\infty} g(r).\exp(-2i\pi.\nu.r.\cos(\theta_o-\varphi)).dr \end{aligned} \tag{75}$$

or

$$FT_2(\tilde{g}_{\theta_o})(\nu,\varphi) = FT_1(g)(\nu.\cos(\theta_o-\varphi)) \tag{76}$$

This equation means that the 2-D FT of a 1-D function placed along a central line with an angle θ_o is equal to the repetition of the 1-D FT of this function, parallely to the line with an angle θ_o. Thus, equation (74) can be written, using the notation of the repetition

$$FT_2(\tilde{g}_{\theta_o}) = \widehat{(FT_1(g))}_{\theta_o} \tag{77}$$

The relation is still good for the inverse FT.

Equation (77) shows clearly the direct relation between the notions of extension and repetition. Indeed, it is a consequence of the fundamental property of the Dirac's distribution : the FT of δ is equal to the unit function.

The four notions defined in this Appendix can be similarly expressed in three dimensions. The equations are quite similar using cylindrical coordinates instead of polar coordinates and g(r,z) and f(r,z) instead of g(r) and f(r).

The "3-D extension" and "3-D repetition" of a 2-D function are still called "extension" and "repetition" without mention of the final dimension. The properties of the superposition and of the 3-D FT of a 2-D function are identical to those in two dimensions.

APPENDIX II

Consider the superposition of the repetitions (cf. Appendix I) of the central slices of the projection picture g of f. Then, one obtains a new 2-D function called "back-projected" function of g and denoted $So\hat{g}_\theta$

$$(So\hat{g}_\theta)(R,\psi) = \int_0^\pi \hat{g}_\theta(R,\psi).d\theta \tag{78}$$

and, using (73)

$$(So\hat{g}_\theta)(R,\psi) = \int_0^\pi \int_{-\infty}^{+\infty} g(r,\theta).\delta(r-R.\cos(\psi-\theta)).dr.d\theta \tag{79}$$

$$(So\hat{g}_\theta)(R,\psi) = \int_0^\pi \int_{-\infty}^{+\infty} g(r,\theta).\frac{1}{|r|}.\delta(r-R.\cos(\psi-\theta)).|r|.dr.d\theta \tag{80}$$

$$(So\hat{g}_\theta)(R,\psi) = \int_0^\pi \int_{-\infty}^{+\infty} g(r,\theta).\delta(r^2-r.R.\cos(\psi-\theta)).|r|.dr.d\theta \tag{81}$$

Equation (81) shows that the back-projected function of g can be obtained by a weighted integration along the circle

$$r.R.\cos(\psi-\theta) = 0 \tag{82}$$

with (r,θ) variable and (R,ψ) fixed. According to (80), it is obvious that the distribution δ of the equation (81) has a density $1/|r|$ with respect to r and $1/|R|$ with respect to r and θ. As the right side of the equations (81) and (13) are identical, it is a proof for equation (12), which is the same equation as (81). Therefore it is a proof of the rigorous equivalence of the back-projection and the circularization techniques. Moreover, it can be shown that the equation (79) allows to obtain the equations (7), (8), (9), (14), (15) directly. Using the property of the FT of the unit function of being equal to δ, and replacing h by Sog_θ in the equation (79) it follows that

$$h(R,\psi) = \int_0^\pi \int_{-\infty}^{+\infty} g(r,\theta).\left(\int_{-\infty}^{+\infty} \exp(-2i\pi.\nu.(r-R.\cos(\psi-\theta))).d\nu\right).dr.d\theta \tag{83}$$

$$h(R,\psi) = \int_0^\pi \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} g(r,\theta).\exp(-2i\pi.\nu.r).dr\right).\exp(2i\pi.\nu.R.\cos(\psi-\theta).d\nu.d\theta \tag{84}$$

The integral between brackets is equal to $FT_1(g_\theta)(\nu)$ and, due to the Central Slice Theorem, it is also equal to $FT_2(f)(\nu,\theta)$. Thus

$$h(R,\psi) = IFT_2\left(\frac{1}{|\nu|}.FT_2(f)\right)(R,\psi) \tag{85}$$

or

$$h = \frac{1}{|r|} * f \tag{86}$$

A straightforward practical technique can be deduced from these equations : the events of the sequential acquisition are placed to their respective positions to form the picture $g/|r|$, then the circularization in the picture g is performed and the deconvolution by $|r|$ is done with an appropriate filter in the

real or in the Fourier domains. The way the equations (83) to (86) have been obtained gives an easy way to prove the fundamental formulas of the 2-D reconstruction. Though the proofs are not so simple, the notions of the Appendix I will be preferred to clearly show the fundamental principles of the 2-D reconstruction.

DISCUSSION:

For the discussion Todd-Pokropek assisted Berche by acting as interpreter.

Goris asked why the centers of the reconstructed images of the liver were hotter than the surrounding regions, since one would have expected them to appear colder due to attenuation. Berche replied that Wiener filtering had the effect of reducing the edges and bringing out the central region.

Berche told Oppenheim that only one slice of the object was reconstructed at a time, rather than the entire three-dimensional object. Oppenheim asked how superimposed out-of-plane information was removed from an oblique slice. Berche responded that this was handled by the method of direct projection in the Fourier domain. Berche informed Erbsmann that the resolution of the technique was approximately 2 to 2.5 cm.

Metz asked if the Wiener filter was a filter which minimized some squared difference criterion, and Berche said that it was, with reprocessing of white noise. Metz wondered how the filter was chosen when with clinical data one does not know what one is trying to fit, and Berche responded that the filter was chosen empirically from experimental results.

Rogers asked why data was collected in list mode rather than fast frame mode, and Berche answered that this avoids the necessity to reconstitute frames as separate projections, so that the process is continuous rather than discrete. Berche informed Soussaline that slice thickness is derived from that value on the z-axis for which photons are kept, and is a purely statistical criterion. The slice is infinitely thin for the frontal or sagittal orientation.

Kirch noted that the subject was far from the collimator, and Berche said the chair would have to be redesigned to reduce this distance. Weisner was told that the axis of rotation of the chair was aligned with the detector face with a plumb line. The alignment did not appear to be very critical.

EMISSION TOMOGRAPHY WITH TIME-CODED APERTURES

Kenneth F. Koral, Ph.D., Glenn F. Knoll, Ph.D. and W. Leslie Rogers, Ph.D.
University of Michigan Medical Center, Ann Arbor, Michigan 48109

ABSTRACT

Two coded aperture geometries are under investigation at the University of Michigan for in vivo imaging of single-photon-emitting radionuclides. The first is a flat plate aperture used in conjunction with an Anger camera detector while the second is a ring aperture employing a concentric ring of individual sodium iodide detectors. In each case, the aperture transmission is time modulated in a pseudorandom sequence; in the first instance by translation of the plate, and in the second by rotation of the ring.

For purposes of tomographic reconstruction, use of a time modulated code is significant in that the temporal correlation coefficients relating the observed count rate and the aperture transmission are directly proportional to projection elements or ray sums.

In the flat plate geometry, the source distribution is incompletely sampled in angle and reconstructions are made as longitudinal tomograms. Previously, these have been generated by additive back-projection. More recently, matrix inversion, image stripping and ART have been employed to eliminate the out-of-focus background from inferior and superior planes. Best results to date have been obtained with a true 3-dimensional form of ART. This algorithm produces qualitatively correct images of phantoms, and tomograms of patient thyroids which are consistent with conventional radionuclide images and ultrasound evaluation.

By virtue of its circular geometry, the ring aperture is capable of yielding transverse section images using conventional Fourier or filtered back-projection techniques. Additive back-projection images of simple source distributions have been obtained with an early prototype equipped with three detector elements. A full detector ring of over 60 detectors is now under construction.

INTRODUCTION

Application of coded apertures to imaging distributions of gamma-ray emitting radiotracers in nuclear medicine has been studied by a number of investigators for over 5 years (1). Coded apertures exhibit improved resolution, single-to-noise ratio, and field of view compared to conventional apertures for a number of imaging situations. One of the key features offered by coded apertures in the nuclear medicine application is a tomographic response, that is, in-focus images may be reconstructed corresponding to a given depth in the

object volume. In most reconstruction methods employed to date, however, out-of focus information from overlying and underlying planes remains superimposed on the in-focus image giving reduced contrast and, in some instances, generating artifacts. The above problem has severly limited the usefulness of coded aperture imaging in nuclear medicine. This investigation was undertaken specifically to explore methods for removing the out-of-focus information from these images. The scope was restricted to image reconstruction for a pseudorandom time-modulated aperture which has been developed at the University of Michigan.

This aperture and its use in gamma-ray imaging is described more fully elsewhere (2,3), however, the essential features of the aperture geometry are illustrated in Figure 1 for a one-dimensional case. The lead aperture plate of width W is divided into L segments of width w whose transmittance for the source radiation is 1 or 0. Counts are recorded in the detector as a function of position as shown in Figure 1 which illustrates the case for one of the source volume elements in the field-of-view.

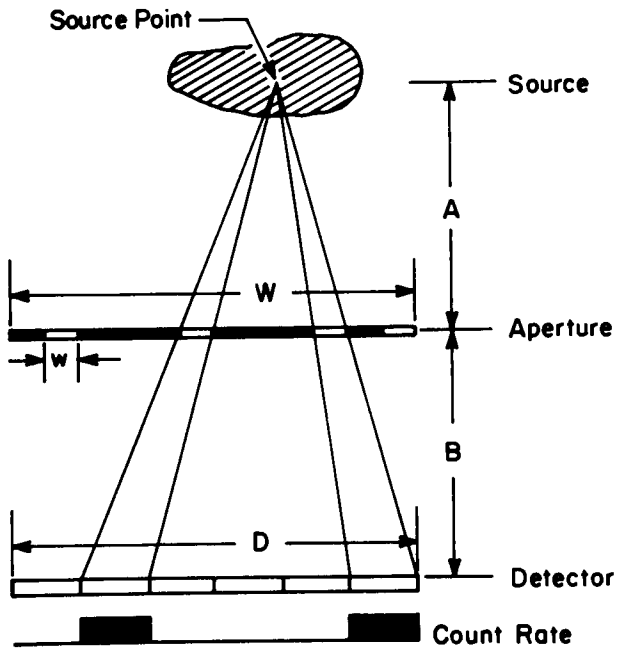


Fig. 1. Illustration of one-dimensional coded aperture with multi-element detector. During a different time interval, another set of open elements will be used and the same source point will give rise to a different count rate distribution in the detector.

Total measurement of time is divided into L intervals. For each interval the aperture code pattern is changed and counts are subsequently recorded for each of the detector elements. Raw data consists then of the counts in the jth detector element during the νth time interval $C_{j\nu}$, and the known transmittance of the ith aperture segment during the νth time interval, $T_{i\nu}$. In this work, the sequence of transmittances is generated by cyclically permuting certain pseudorandom binary strings (4,5) in which each aperture segment describes the same binary sequence but out of phase with all other segments. This means that a sequence of length L may be used to modulate a total of L segments of the aperture. These pseudorandom strings exhibit peaked autocorrelation functions with flat side lobes. This property of the code makes it possible to calculate the integral source strength along a line seen by any detector element through any aperture segment by correlating the observed count rate with the known aperture segment transmittance over the total measurement time as shown in Equation 1.

$$Q_{ij} = \sum_{\nu=1}^{L} C_{\nu j} \; T_{i+\nu} - m \sum_{\nu=1}^{L} C_{\nu j} \qquad (1)$$

Q_{ij} is just the projection element or ray integral along rays directed from the jth detector element through the ith aperture segment. The constant m in Equation 1 is the side-lobe-to-peak ratio of the code sequence employed, and subtraction of this term completely removes the side lobes from the point response function. This type of coded aperture has been experimentally realized in two different geometries. The initial approach was to use a plane rectangular aperture with an Anger camera as the detector. The bulk of the work reported here deals with image reconstruction techniques appropriate to this situation. More recently, Knoll and Williams have described a time-coded ring aperture utilizing

a ring of discrete detectors (6). Image reconstruction in ring geometry poses a different set of problems from the flat-plate geometry and will be discussed briefly in a later section.

TOMOGRAPHIC RECONSTRUCTION - PLANE APERTURE

Basic Considerations

Figure 2 illustrates the geometry for a hypothetical aperture using a code of length L = 25 arranged in a 5 x 5 array. In practice, the transmission of the aperture is modulated by translating a lead code plate in the aperture plane. The coordinate information from the detector is digitized and the coded image corresponding to each of the L intervals is stored as a 64 x 64 matrix on magnetic disk. Calculation of correlation coefficients by Equation 1 yields an array of 4,096 x L projection elements. It is this collection of projection elements which characterizes the image reconstruction problem, and the fact that they are derived from a time-coded aperture is not relevent except as the technique influences the geometry and signal-to-noise ratio of the projection elements. A point in the object volume will be described by a subset of these projections contained in a pyramid shaped volume as sketched in Figure 2. The angular range, Θ, which these projections cover depends on the distance from the source point to the aperture, A, on the size of the aperture and detector, and on the lateral position of the object point. When the aperture is smaller than the detector, Θ for distant on-axis points will be limited by the aperture size. As the point approaches the aperture, Θ increases and the aperture shadow cast by the point expands until it exceeds the detector size. Now Θ is determined by the size of the detector. Also, as the object point approaches the aperture, the sampling interval, $\Delta\Theta$, between adjacent projection elements increases. As the object point is moved laterally such that the detector intercepts fewer rays, Θ is decreased.

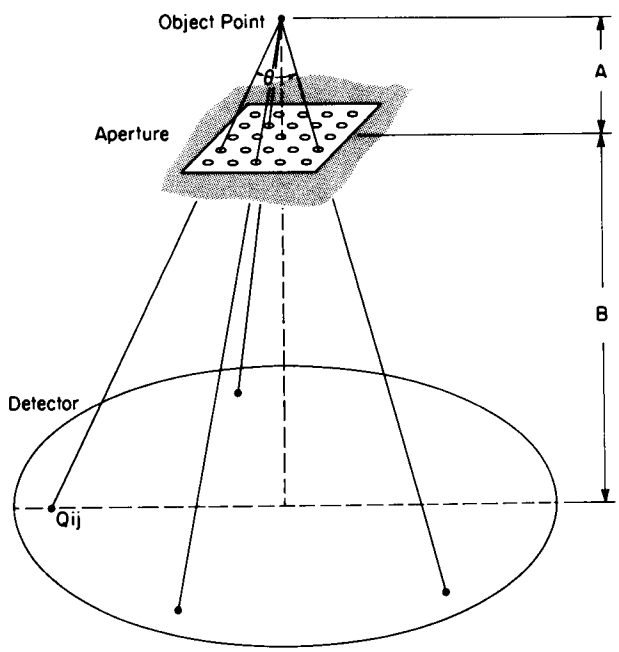


Fig. 2. Two-dimensional aperture shown approximately to scale of experimental aperture and Anger camera. A few of the extreme projection elements, Q_{ij} which intersect an object point a distance A from the aperture are illustrated. The angular range spanned by these projection elements is indicated by Θ.

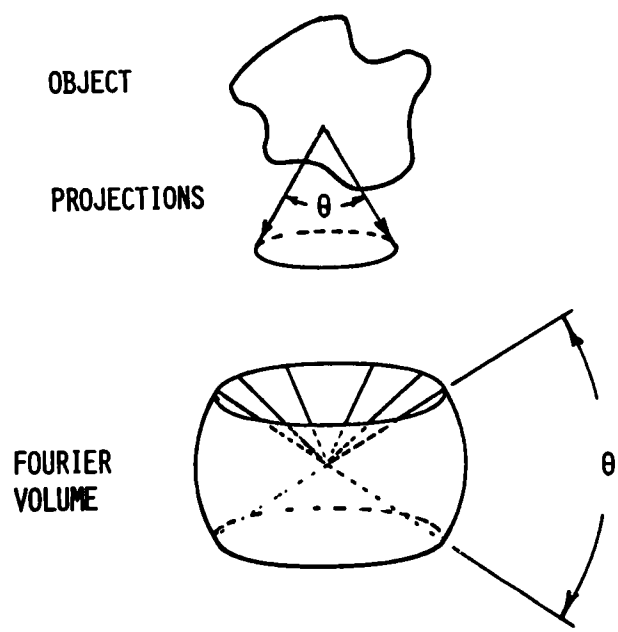


Fig. 3. Relationship between set of conical projections and region of Fourier volume sampled.

Given this subset of projection elements for rays intersecting the illustrated object point and spanning an angular range Θ, several methods may be employed to reconstruct the intensity of this point. Before describing these, however, limitations posed by this geometry will be examined.

Principal difficulty arises from the limited angular range of Θ which determines the region of the object's Fourier spectrum that is sampled. Considering the projection theorem in three dimensions shows that a conical set of rays will sample the Fourier volume over a spherical region with a conical void as illustrated in Figure 3. This means that any attempt to unambiguously reconstruct the object from noisy data is bound to meet with limited success, and,in general, resolution perpendicular to the aperture is degraded relative to that in planes parallel to the aperture. It should be noted that the result will be object and orientation dependent as governed by the significance of the unsampled region of the object spectrum and the success of the reconstruction method in recovering this unsampled data. This type of sampling is peculiar to a time-coded planar aperture which yields information on individual projection elements. Stationary planar coded apertures such as the Fresnel zone plate only sample the Fourier volume on a surface (7).

Combined with the incomplete sampling problem is the fact that measured projection-element strength exhibits a strong angular dependence arising from variations in solid angle as given in Equation 2:

$$\Omega_{ij} = \alpha \mathrm{Cos}^3\Theta_{ij} \sqrt{(w-t\ \tan\Theta_{ij})^3\ \ (w+t\ \tan\Theta_{ij})}\ . \qquad (2)$$

The term under the radical sign is an approximation for the open solid angle for round holes of diameter w in an aperture plate of thickness t. Attenuation is a factor of five between 0 and 40°.

In choosing a method to reconstruct the coded-aperture images it is necessary to consider the preceeding difficulties and document the sensitivity of the reconstruction technique to them. Since the spatial resolution in the direction normal to the aperture is worse than in planes parallel to the aperture, all reconstructions are longitudinal planes parallel to the aperture. Three approaches were investigated for image reconstruction:

a) matrix inversion

b) image stripping

c) ART

In the sections that follow, each of these are described and representative results are shown.

Matrix Inversion

This method, described by Chang and Macdonald (8) is illustrated in Figure 4. Rays from two point objects on the left are shown intercepting the detector. Image reconstruction by back-projection and superposition of the projection elements is shown to the right. For the centrally located source all rays projected through the aperture intercept the detector and the superposition image, T_j, may be written as the sum of convolutions of the object planes, O_i, with the appropriate transfer functions H_{ij} as shown. The superposition image comes into focus in plane 1 and defocuses to form a scaled image of the aperture array in planes other than the focal plane. One of the experimental problems alluded to above is portrayed by the other source, denoted by an X, which

is located toward the edge of the field of view. Some of the rays projected through the aperture from this source miss the detector and the in-focus image is reduced in intensity while the blur function is truncated.

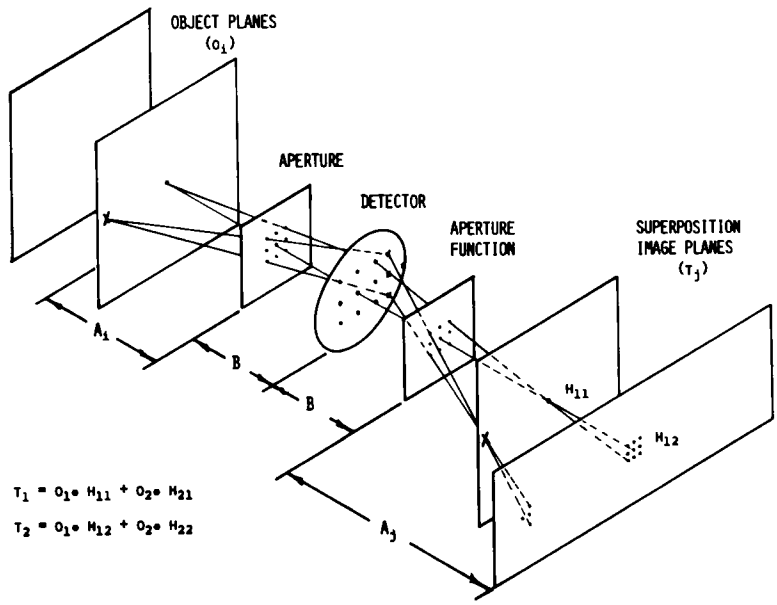


Fig. 4. Illustration of matrix formulation of image reconstruction.

Under the assumption that:

a) the aperture holes can be approximated by δ functions,

b) the object consists of discrete planes,

c) and the system transfer function is invariant under translation, i.e. that all of the aperture elements project onto the detector and variations of solid angle with source translation can be ignored,

the transfer function, following Chang's development may be written as:

$$H_{ij} = \sum_{1}^{L} \delta(\vec{r}-(\frac{\vec{r}A_j}{A_i} + M_{ij}\vec{r}_k)) \tag{3}$$

where

$$M_{ij} = \frac{A_i - A_j}{A_i}$$

determines the scale of the transfer function and A_j/A_i determines the displacement. If the equations for T_i are Fourier transformed, the convolution operation is replaced by a product of the object spectrum with the MTF, and the object spectrum may be solved for by matrix inversion. An inverse transform gives the object distribution.

The determinant involved in the matrix inversion is always 0 at the origin which means that the d.c. level of the result is indeterminant. In all figures presented herein, the images are biased so that the most negative point is 0 with the justification that negative strengths are unphysical. The determinant for the δ-function H's of equation 3 increased steadily for spatial frequencies above 0. When transfer functions were used which accounted for the finite aperture hole size, the determinant exhibited oscillations.

Figures 5 and 6 illustrate reconstructions obtained from simulated data for a noise-free point source and one-dimensional geometry. For this simulation and those that follow, the projections are limited to a plane by the one-dimensional geometry. Scale factors were chosen for the simulation to correspond to the experimental geometry so that the detector is 25 cm wide and 15 cm from the

aperture which is in turn 3.9 cm wide with 3.6 mm aperture elements. Solid angle factors were neglected in both the data simulation and the reconstruction. Figure 5 illustrates the effect of finite aperture hole size. The right-hand column is a perfect reconstruction of a point source in plane 3 and was obtained by using a transfer function for a finite sized aperture hole. If the δ function H_{ij} of Equation 3 is employed, the results are those illustrated in the center column which show an undershoot in plane 1 and significant activity in plane 2. The left-hand column shows the superposition images, T_i. Figure 6

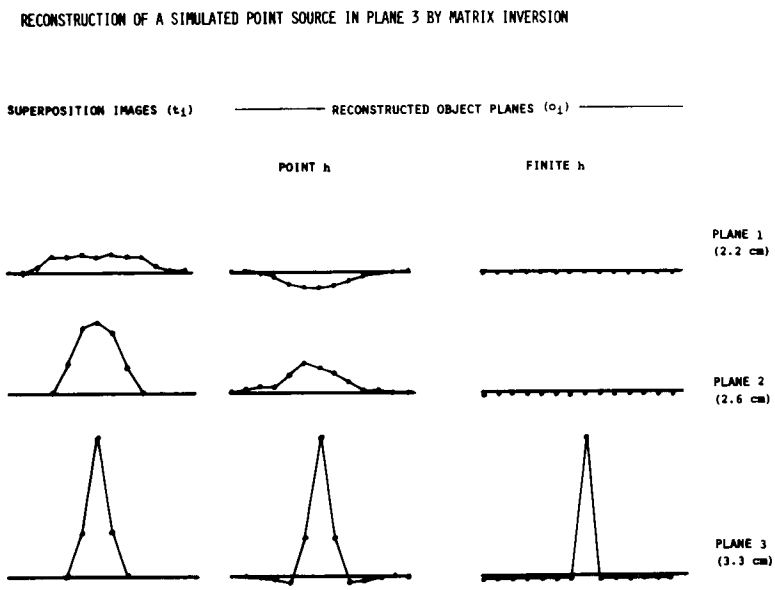


Fig. 5. Illustration of the sensitivity of the reconstruction to the assumed size of the aperture holes for a one-dimensional aperture. Noise-free data was simulated for 3.6 mm aperture holes.

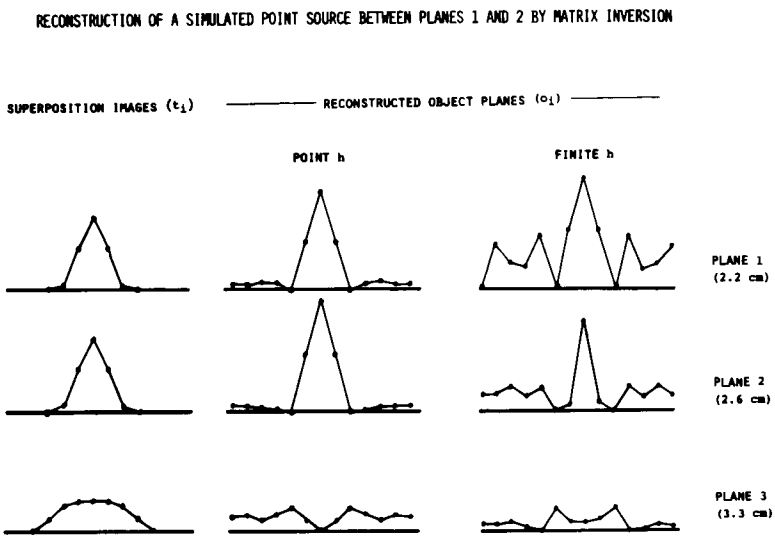


Fig. 6. Illustration of the sensitivity of reconstruction to assumed size of the aperture holes when the source is not confined to one of the object planes. Reconstruction using the actual transfer function with the inconsistent data causes oscillation. As in Figure 5, the simulation is one-dimensional and noise-free.

shows what happens to a point source reconstruction when the source is located halfway between the first two object planes. In this instance, the δ-function H gives better results than the finite H which shows a tendency to oscillate with the inconsistent data. The superposition images here could be considered superior to either of the others in that the image defocusses smoothly and predictably. This inability of the matrix-inversion technique to handle point sources between planes, which is not the case with the other methods to be discussed, forecasts problems in reconstructing volume sources and may reflect a particular sensitivity of matrix inversion to the error involved in replacing integrals over depth by sums over a finite number of planes (equations in Figure 4). Figure 7 illustrates two reconstructions from actual three-dimensional experimental data acquired with an 11 x 11 33%-open coded aperture and Anger camera. Since experimental object distributions are continuous and the data is

noisy, the point H_{ij} of Equation 3 was used for the illustrated reconstructions. In the experimental situation, the effects of solid angle on the response function were not considered. The results exhibit problems which are similar to those in the simulations although not identical.

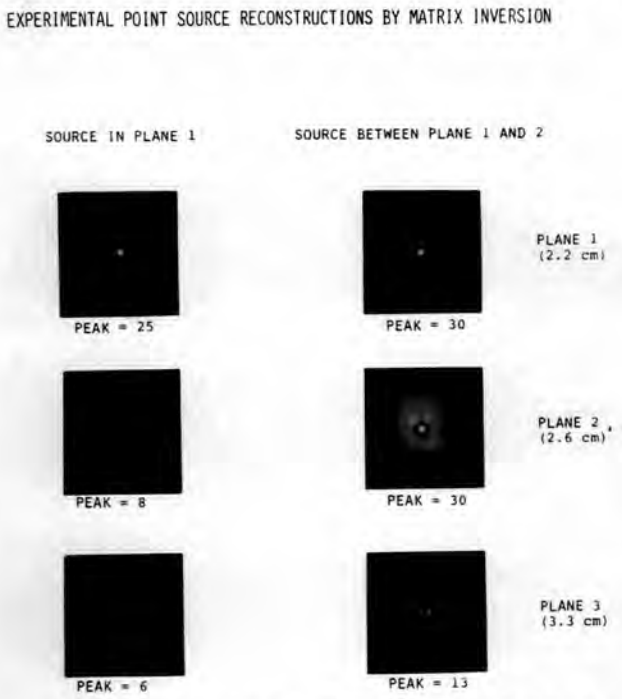


Fig. 7. Application of matrix inversion to experimental point source data obtained with an 11 x 11 coded aperture and Anger camera. A δ-function transfer function was assumed and no solid angle corrections are made. Note tendency to overshoot in plane 2 and the appreciable residual source intensity in planes other than the object plane.

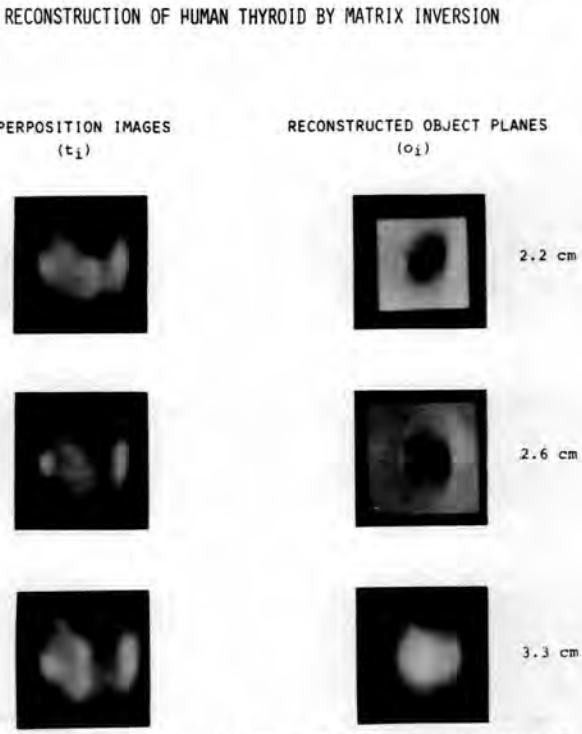


Fig. 8. Image of human thyroid reconstructed by superposition and by matrix inversion. Aperture was located 1 cm from neck. A dose of 5 mCi of 99m technetium was used and imaging time was 6 minutes. Only the furthest plane bears any resemblance to the object.

When a larger volume source is imaged close to the aperture such as was the case for the patient thyroid illustrated in Figure 8, many of the source points do not cast a complete shadow of the aperture on the detector. The thyroid reconstructions now bear little resemblance to the object. This rather complete failure led us to look for a reconstruction technique that operates in the spatial domain with the hope that finite geometry would cause fewer problems and that the corrections could be more readily handled.

Image Stripping

This method is an attempt to build up the object distribution from features extracted from the superposition images. All of the superposition image planes are searched for a maximum image element. A preset fraction of this image point intensity is used as a threshold. Everything in the plane which exceeds this threshold is considered an in-focus feature and is both subtracted from the superposition image T_i, and added to the corresponding reconstructed object plane, O_i. The contribution of these features to the other superposition image planes is calculated by convolving the features with the response function, H_{ij}, for each plane, and subtracting the result from the corresponding superposition image, T_j. The resulting superposition images are now searched for a new feature and the process is repeated. This continues until the superposition images are reduced to zero.

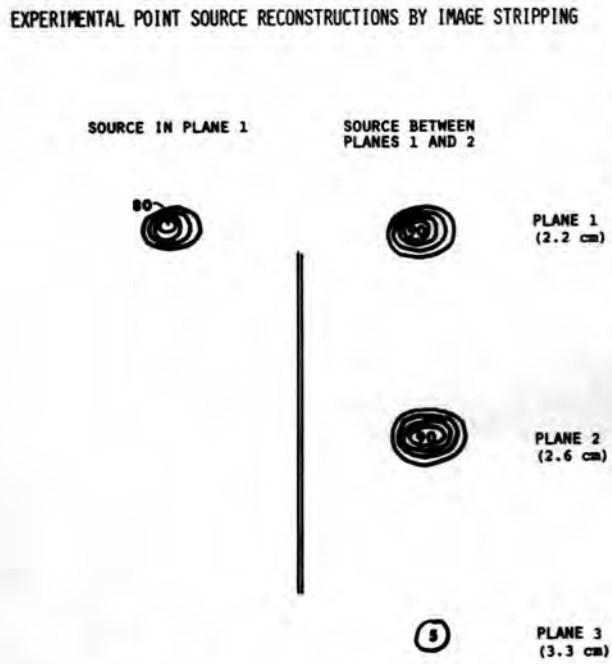


Fig. 9. Results of image stripping for point source reconstruction from experimental data. The transfer function employed for both sets of images was for a 3.6 mm diameter aperture hole. Contour levels for the source in plane 1 are at 5, 40, 60 and 80 counts while for the source between planes the levels are at 5, 30, 60 and 90 counts.

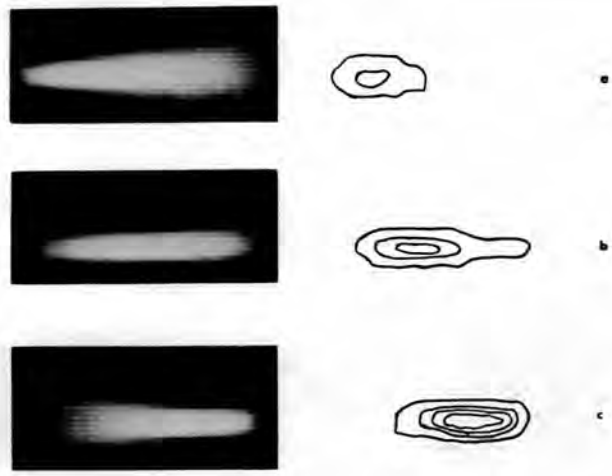


Fig. 10. Comparison of superposition images (left) to reconstruction by image stripping for an experimental line source sloping away from the aperture at about 20°. Contour lines are at 2, 25, 50, and 75 percent of maximum strength. The planes are located at 2.2 cm (a), 2.65 cm (b), and 3.3 cm (c) from the aperture.

This process gives perfect results with simulated point sources and the results obtained for the experimental point sources described above are shown in Figure 9. These results also are virtually perfect both for a source in plane 1, and for a source between planes 1 and 2. Figure 10 shows the reconstruction of an experimental line source which slopes away from the aperture at about 20°. Ideally the sections through this source would be ellipses with approximately a 3:1 ratio between major and minor axes. The stripped images represented by contour lines at 2, 25, 50, and 75 percent of peak strength are quite good in this respect. The difficulty observed here is that the rear plane (c) contains a much greater portion of the activity than the other plane. This is a direct consequence of portions of the aperture shadow falling beyond the detector boundary for near regions of the source, especially those toward the edge of the field. The apparent image intensity in these regions is reduced because fewer rays are intercepted by the detector as shown in Figure 4 for the X-shaped source. The problem is further complicated because the blur function changes shape with source position and improper subtraction of out-of-focus information results. Figure 11 illustrates the problem directly with simulated data for a one-dimensional aperture 3.9 cm long. The source distribution consists of two parallel line sources in planes 1 and 3. The source in plane 3 is 5.2 cm long and is located 1.1 cm behind a 3.5 cm long source in plane 1. This is a relatively difficult source distribution to reconstruct since the projections only cover a range of 61° for the center of the line in plane 3. The angular range subtended by projections at the end of the line depends on the detector size assumed and is 28° for a 25 cm long detector and 42° for one of infinite length. With an infinite detector, the reconstruction is moderately good considering the nature of the object and the angular range of the projections. On the other

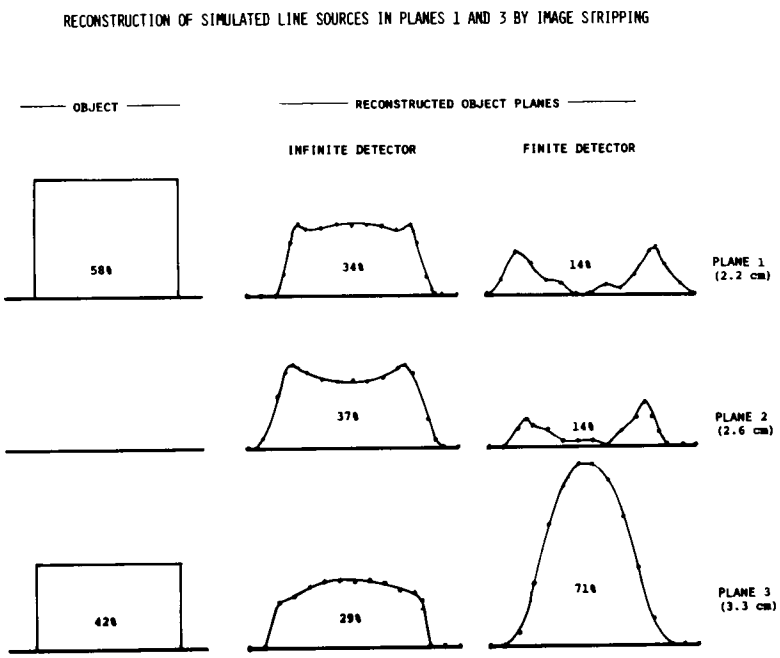


Fig. 11. Effect of detector size on reconstruction by image stripping. Aperture and detector are one-dimensional. Simulated line sources are positioned one behind the other. Plane 1 contains a 3.5 cm long source and plane 3 contains a 5.3 cm long source. These sources are portrayed the same length because they are scaled in image elements whose dimension varies from plane to plane.

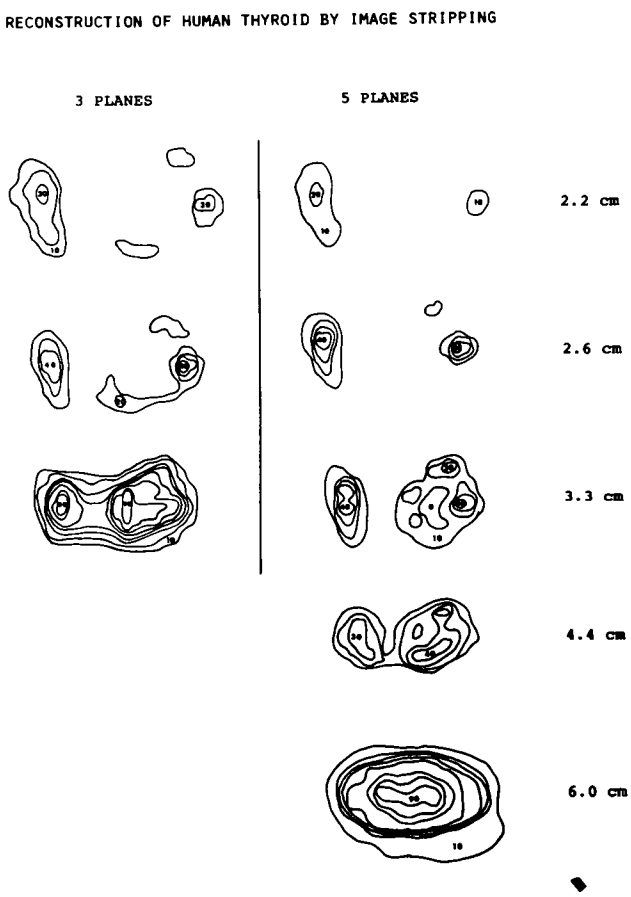


Fig. 12. Reconstruction by image stripping with the same thyroid data used for Figure 8. Images have been reversed from left to right. The tendency for increased source strength to be assigned to the most distant planes is observed regardless of the number of planes reconstructed.

hand, when the detector is finite, the reconstruction fails entirely. Figure 12 shows the result of image stripping for the same thyroid portrayed in Figure 8. With 3 planes, concentration of activity in the back-plane is again observed, and with 5 planes the activity migrates even further back. Although this problem could be partially solved by correcting the superposition image intensities and employing a spatially variant transfer function, the method would be inexact and awkward. For this reason ART was investigated as an alternate approach.

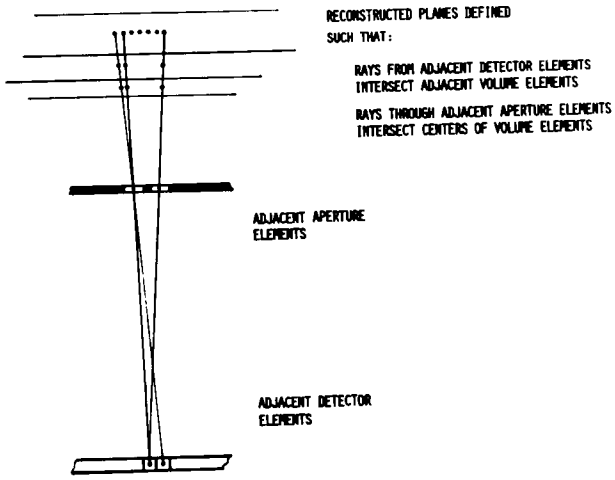


Fig. 13. Definition of reconstruction geometry to avoid interpolation. To maintain integer relationship, planes are not uniformly spaced in depth and the scale factor in terms of millimeters/image element varies from plane to plane.

ART

Iterative reconstruction for the general type of geometry was shown by Colsher (9) to give fairly reasonable results in both noise-free and noisy simulations of spherical objects. It is also possible, as will be described, to precorrect the data for variations in solid angle. Since no assumptions need be made concerning invariance of the transfer function, the reconstruction problem is greatly simplified.

Again, the reconstruction is fully three-dimensional since the rays are not confined to a single plane. The object volume to be reconstructed is not divided into uniform cubes as usual, but is divided so that all rays intercept only the centers of volume elements. This is illustrated in Figure 13. To satisfy this condition only certain object planes may be reconstructed, and each object plane will have a different scale factor in terms of millimeters per image element. From Figure 13 this scale factor is

$$S = \frac{A_n}{B} \times d \tag{4}$$

Here A_n is the distance from the nth object plane to the aperture, B is the aperture-detector separation, and d is the distance between adjacent detector elements. The values of A_n which satisfy the integer relationship are given as

$$A_n = \frac{B\ell}{nd-\ell} \tag{5}$$

The distance between adjacent aperture elements is ℓ and n is restricted to integer values. Defining the object volume in this manner achieves two things. First, the computation time is reduced since integer arithmetic is used without need for interpolation. Second, the distance between reconstructed planes increases as distance from the aperture increases, but in a manner which is consistent with the reduced depth resolution expected for more distant object planes where the angular range sampled is smaller.

The measured values for each of the ray sums or projection elements depends upon the solid angle given in Equation 2. Because of the manner in which the volume element is defined, the solid angle factor depends only on Θ_{ij} and not upon the distance from the plane to the aperture. For this reason, the projection elements can be corrected before starting the reconstruction. It should be emphasized that the solid angle correction is necessary to make projection elements at different angles consistent with each other.

The image grid defined by the integer relationship is considerably finer than the image resolution so that the superposition images may be condensed 3:1 in each direction without reducing resolution. This was done for both matrix inversion and image stripping to reduce computer memory requirements and computation time. However, when this was attempted for the ART reconstructions, difficulty was encountered in that the projection elements can't be condensed so that they bear a 1:1 relationship with the condensed images. In order to perform the computation in a mini-computer with a 32k memory images are constructed iteratively from each of 81 subsets of projection elements. Iteration continues for either a preset number of times or until the change in variance from the mean per iteration is less than a preset amount. The images from each subset of data are added together. The ART reconstruction of a simulated point source is essentially perfect while the reconstruction of an experimental point source in plane 1 leaves residual peak intensities of 14% and 7% in planes 2 and 3 respectively.

Figure 14 illustrates ART results obtained for the same one-dimensional aperture and noise-free line sources of Figure 11. For the infinite detector the ratio of source strength in plane 1 to that in plane 3 is nearly the same as

the object, and the distribution between planes 1 and 3 is smooth. Again, because of the reduction in angular range, results for the finite detector are not as good as for the infinite detector, but they are far superior to those obtained by image stripping. Two experimental results are shown in Figure 15 and 16. In each, the superposition images are compared to the ART reconstructions. The letters of Figure 15 are made from 99mTc filled capillary tubing and the iteratively reconstructed images are markedly superior. The thyroid images are reconstructed from the same data as Figures 8 and 13. These reconstructed images correspond well to the ultrasonic finding of a large cystic nodule in the right lobe.

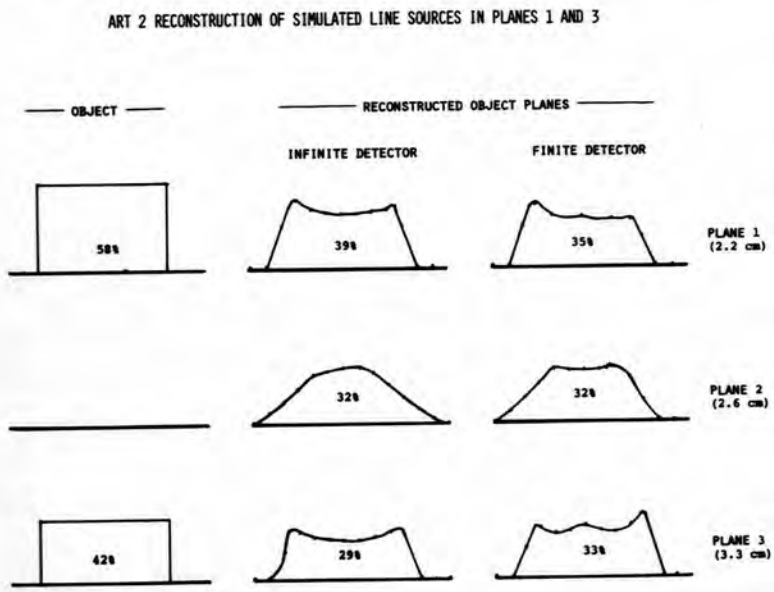


Fig. 14. Reconstruction by ART of same line sources portrayed in Figure 11. Reducing the view angle by using a finite detector degrades the results and emphasizes an asymmetry which appears to reflect the order in which iteration is performed.

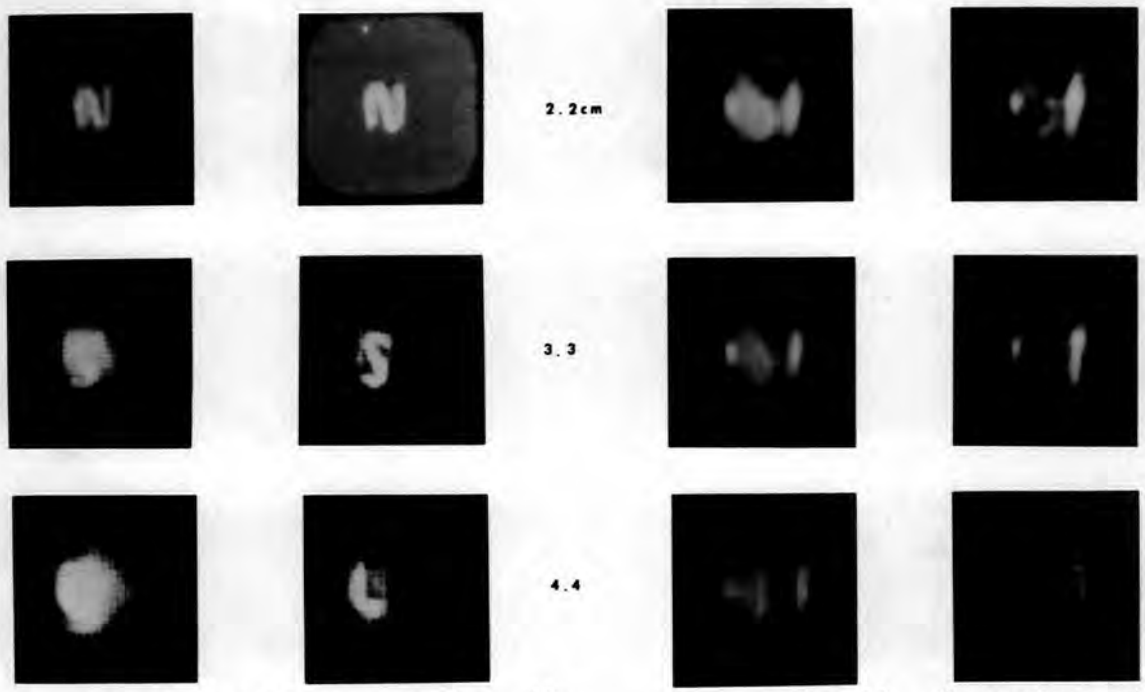


Fig. 15. Letters made from 99m technetium filled tubing. Reconstruction by back-projection and superposition (left) is compared to reconstruction by ART. Data was obtained by using an 11 x 11 coded aperture with Anger camera detector. Some residual out-of-focus background is evident, but the image improvement is substantial.

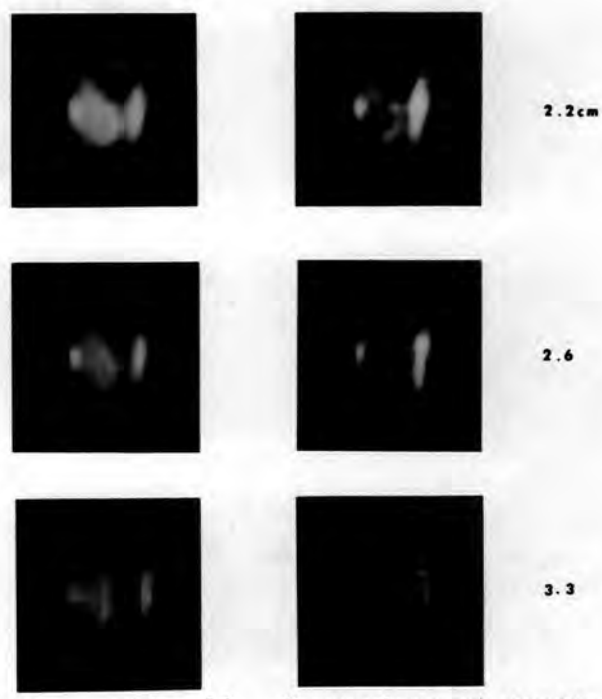


Fig. 16. Comparison of superposition images (left) and ART reconstruction of human thyroid. c.f. Figures 8 and 12. The images are consistent with ultrasound images which show a large cystic nodule in the right lobe of the thyroid. Adding 1 or 2 reconstruction planes produces little change in these images.

Coded Ring Aperture

All of the previously described methods of image reconstruction have been directed toward the case of limited angle, planar aperture, geometry. Although this aperture is readily used with an Anger camera detector, the problem posed by incomplete projections is substantial. This problem is eliminated by using the coded aperture in ring geometry as reported by Knoll and Williams (6). In this scheme, a lead ring with vertical slits arranged in a pseudorandom pattern is used as a coded aperture. Discrete detectors are arranged in a larger ring concentric with the aperture ring and are collimated so that they view a narrow slice in the plane of the aperture ring. The geometry is shown in Figure 17 for three detector elements. Gamma ray transmission of the aperture is modulated by rotating the ring, and just as described earlier, a correlation calculation yields a set of projection elements or ray sums for each detector. Work is now underway to complete the ring of detectors. When this is accomplished, the region within the ring will be fully sampled over an angular range of 360°.

As of this time, experience with reconstruction algorithms for the ring geometry has been confined to superposition, but it is evident that any of the fan beam algorithms should be applicable with only minor modifications. Figure 18 illustrates a superposition reconstruction of a point source and shows the experimental and calculated variance for the measured source strength.

DISCUSSION

Of the three methods investigated for reconstructing images for the plane aperture geometry, ART consistently yields superior results to matrix inversion or stripping. Much of the difference is attributable to being able to readily compensate for the spatially variant response of the imaging system. Matrix inversion and stripping are both attractive methods because all operations are performed on the superposition images. This brings about a significant reduction in the amount of data that must be stored. If four planes are to be reconstructed, this means the input data consists of only 16,000 numbers rather than the 495,000 individual projection elements which are obtained in our experimental arrangement. Because the efficiency and response function is strongly translation dependent, however, this compression of the data proves to be a disadvantage. The principal disadvantage of ART is the length required for image reconstruction. Over two hours of mini-computer time are required for three iterations on four planes. There is little question that this reconstruction time could be reduced by a factor of ten or more using relatively inexpensive special purpose hardware so that computing time need not be regarded as an insuperable obstacle to iterative reconstruction.

Another approach which has some elements in common with the image stripping method is reported by Atkins in these proceedings(10). This method involves iteration on the superposition images rather than the individual projection elements and has been applied for reconstructing images from unrotated positron cameras. With the proper compensation for coded aperture geometry this might be an attractive technique since it is an iterative method yet still has the advantage of reduced data storage requirements.

Since the coded ring aperture samples the object over a full 360° it is expected that the reconstructions obtained will be superior to those for the flat plate aperture geometry. The present configuration of the ring aperture has the disadvantage of only imaging a single slice. The planar aperture, on the other hand, images a volume. It also employs a standard Anger camera as detector. As such it is readily portable and can be used with seriously ill patients at bedside. It is a question of whether the images are diagnostically useful in spite

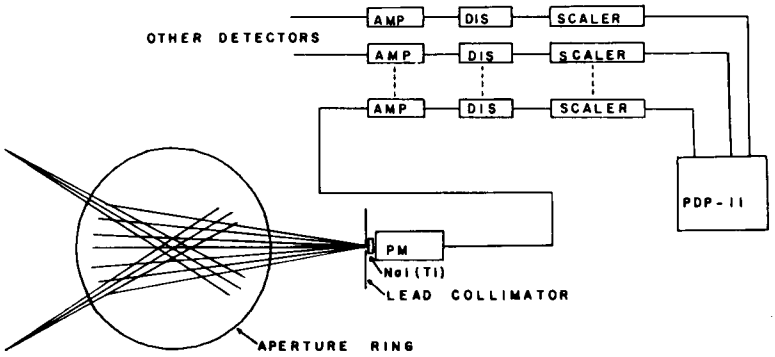


Fig. 17. Illustration of coded ring aperture geometry showing three detectors. Aperture transmission is modulated by rotating the ring. The detectors are collimated in the plane of the ring and the decoded data consists of ray integrals for the fan beams illustrated for each detector.

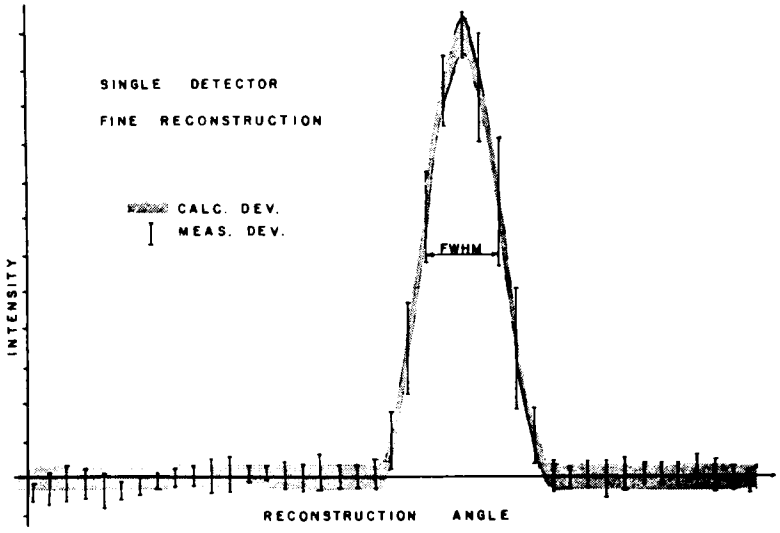


Fig. 18. Point source reconstruction for coded ring aperture by superposition back-projection. The FWHM value shown is that calculated from experimental geometry. Error bars indicate experimentally measured deviations for 8 independent measurements and shaded region indicates predicted deviations.

of their recognized shortcomings. To answer this will require additional study.

REFERENCES

1. W.L. Rogers: Coded Aperture Imaging in Nuclear Medicine; Review and Update, Proceedings ERDA X and Gamma Ray Symposium. Ann Arbor, Michigan May 19-21, 1976. (CONF 760539).

2. K.F. Koral, W.L. Rogers, G.F. Knoll: Digital Tomographic Imaging with Time-Modulated Pseudorandom Aperture and an Anger Camera, J Nucl Med 16: 402-413, 1975.

3. R.S. May: Gamma Ray Image Formation by Stochastic Time Modulation, Ph.D. Thesis, University of Michigan, 1974.

4. F. Hossfeld, R. Amadori: On Pseudorandom and Markov Sequences Optimizing Correlation Time-of-Flight Spectrometry, Kernforschungsanlage Julich Report Jul-684-FF, 1970.

5. G. Wilhelmi: Datenverarbeitszentrale. Eine Ubersicht uber Differenzmengen mit Tabellen. Kernforschungszentrum Karlsruhe Report KFK 1177, 1970.

6. G.F. Knoll, J.J. Williams: Application of a Ring Pseudorandom Aperture for Transverse Section Tomography, IEEE Transactions Nuclear Science NS-24 #1 581-583, (1977).

7. R.G. Pridham, A.G. Lindgren: Imaging of Volume Objects by Shadow Casting, Submitted to IEEE Transactions on Nuclear Science.

8. L.T. Chang, B. Macdonald, V. Perez-Mendez: Three-Dimensional Image Reconstructions from Axial Tomography. Proceedings Meeting on Image Processing for 2-D and 3-D Reconstructions from Projections, Stanford, California, August 4-7, 1975.

9. J.D. Colsher: Iterative Three-Dimensional Image Reconstructions from Projections: Applications in Electron Microscopy, Ph.D. Thesis, University of Pennsylvania, December 21, 1976. Also as UCRL-52170, Lawrence Livermore Laboratory, University of California, Livermore, California 94550.

10. F. Atkins, G. Muehllehner, P.V. Harper: Positron Emission Tomography Using Large Area Detectors, Proceedings 5th International Conference on Information Processing in Medical Imaging. June 27-July 1, Nashville, Tennessee 1974.

DISCUSSION:

Rogers informed Kirch that the dog heart images were corrected for magnification by scaling for source distance and that 40 pinholes were open at one time. An aperture under construction would have fewer holes and a signal-to-noise ratio equivalent to 12% transmission. Muehllehner wondered why the counts drift to one image in the image stripping technique. Rogers thought this might be due to an error in the solid angle correction or may be a property of ad hoc stripping of features.

Rogers told Weisner that ring width changes in order to obtain an integer relationship between voxels, and that they have not reconstructed anything deeper than 9 cm. Also, the new cardiac aperture will be larger than the thyroid aperture but will subtend the same range of angles, so that cardiac and thyroid reconstructions should be of similar quality. Brill asked whether circular geometry would overcome field-of-view limitations of planar geometry, and Rogers responded that it will, since with circular geometry the data set is complete.

Asked by Kirch about the relative sensitivity of the coded aperture over the heart, Rogers answered that imaging time would be reduced by a factor of 2 or 3 compared with the scanning pinhole with the same field of view. The scan of a dog heart with 2mCi of thallium took 20 minutes, using the thyroid aperture. Rogers told Pizer that the FORTRAN program for image reconstruction now runs in six hours, but assembly language programming should reduce this time significantly.

Erbsmann asked whether a coded aperture system yielded more information than other tomographic systems and Rogers replied that the coded aperture system should image hot, low contrast regions with greater efficiency than the other systems. Oppenheim questioned whether there was an advantage to using a coded aperture over a collimator which was chosen so that the object completely filled the detector, and Rogers replied that if the object really did fill the detector, this would be the optimal imaging situation for that collimator, and a coded aperture would probably not offer an advantage unless one desired tomographic imaging.

Twieg inquired about gating of the cardiac cycle for cardiac imaging, and Rogers said that this might be implemented at some time.

THE EXPONENTIAL CONVOLUTION ALGORITHM FOR EMISSION COMPUTED AXIAL TOMOGRAPHY*

O. J. Tretiak and P. Delaney
Drexel University
Philadelphia, Pennsylvania

ABSTRACT

The Exponential Convolution Algorithm (ECA) is a concept for correcting the effects of attenuation in radionuclide emission computed tomography. The algorithm is developed and some tests are made with simulated data. The performance of ECA is compared with two ad-hoc attenuation correction procedures.

INTRODUCTION

Since the count rate observed by a collimator or a gamma camera depends not only on the total amount of radioactive material but also on the attenuation of the intervening matter, the reconstruction algorithms for computed axial tomography (CAT) theory using transmitted x-rays are not entirely applicable to radionuclide (emission) computed axial tomography (ECAT). The algorithm used for ECAT must, in some way, correct for the attenuation. Several such algorithms have been proposed. Kuhl, et al. (1) used an iterative procedure that ignored attenuation, followed by a multiplicative correction on the computed tomogram. Budinger and Gullberg (2) made a number of suggestions. One was to use an iterative algorithm which models the attenuation effects in the successive evaluations of the scan data. Kay and Keyes (3) also described an ad-hoc correction procedure. Kuhl, et al. (4) developed a method for correcting the data prior to an attenuation-free reconstruction algorithm. Perhaps the most comprehensive is the method due to Walters, et al. (5). A convolution method is used for the reconstruction and the differences between the true data and simulated data computed from the reconstruction is used to refine the reconstruction. This is a marriage of the convolutional reconstruction procedure first published by Ramachandran and Lakshminarayanan (6) and of the attenuation modified iterative process suggested by Budinger and Gullberg.

The exponential convolution algorithm (ECA) to be described below is based on an analytical inversion procedure. Under condition in which it is valid it promises to perfectly compensate for the effects of attenuation. Its computational speed is comparable to that of the usual convolutional algorithm.

*This work was supported in part by contract NIH-NO1-CB-53859, and in part by NIH grant RR-15 to the Medical School Computer Facility of the University of Pennsylvania.

The ECA is based on the following idealized situation:

a) The activity distribution to be reconstructed is defined on a plane section through a body and is given by f(x,y).

b) The body, on this plane section, has uniform and known attenuation coefficient μ, the shape of the section is known, and it is convex.

c) The data is a family of line integrals of the activity weighted by the attenuation. Referring to Figure 1,

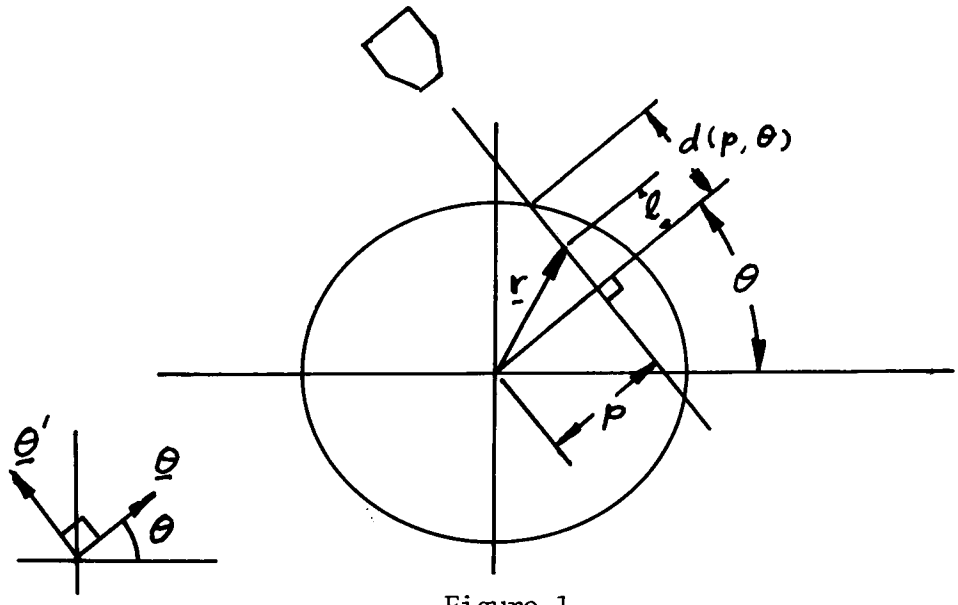


Figure 1

Section Geometry

$$g(p,\Theta) \triangleq \int d\ell \; f(\ell)\alpha(\ell) \tag{1}$$

$$\alpha(\ell) \triangleq \exp\left(-\mu(d(p,\Theta) - \ell)\right) \tag{2}$$

The above equations are written in an informal notation. We represent the (x,y) coordinate pair by the vector $\underline{r}$; $\underline{\Theta}$, is a unit vector in the Θ direction

$$\underline{\Theta} \triangleq (\cos\Theta, \sin\Theta)$$

while $\underline{\Theta}'$ is a unit vector orthogonal to $\underline{\Theta}$,

$$\underline{\Theta}' = (-\sin\Theta, \cos\Theta).$$

In terms of these, the model equations are

$$g(p,\Theta) = \int d\underline{r} \; f(\underline{r}) \; \delta\,(p - \underline{r}\cdot\underline{\Theta}) \; \exp\left(-\mu\left(d(p,\underline{\Theta}) - \underline{r}\cdot\underline{\Theta}'\right)\right) \tag{3}$$

The line integral, $f(\ell)$, is a region of singularity of the delta function $\delta(p - \underline{r}\cdot\underline{\Theta})$. The dot product is used in the usual way

$$\underline{r}\cdot\underline{\Theta} = x\cos\Theta + y\sin\Theta \tag{4}$$

ECA is based on the following manipulations:

1) Since the section shape is known, we know for each (Θ,p) the value of the distance function $d(p,\Theta)$, so that we may form the preweighted projections.

$$g_1(p,\Theta) \triangleq g(p,\Theta)\exp(\mu d(p,\Theta)) = \int d\underline{r}\ f(\underline{r})\delta(p-\underline{r}\cdot\underline{\Theta})\ \exp\mu(\ \underline{r}\cdot\underline{\Theta}') \tag{5}$$

2) We now form a new function by convolving $g_1(p,\Theta)$ with a weighting function $h(p)$

$$g_2(p,\Theta) = \int ds\ h(s)\ g_1(p-s,\Theta) \tag{6}$$

3) We finally form a so-called exponential backprojection at point $\underline{\rho}$

$$f_1(\underline{\rho}) = \int_0^{2\pi} d\Theta\ g_2(\underline{\rho}\cdot\underline{\Theta},\Theta)\ \exp\mu(-\underline{\rho}\cdot\underline{\Theta}') \tag{7}$$

The merit of these steps is seen if all of these steps are combined, because then we see that

$$f_1(\underline{\rho}) = \int d\underline{r} \int ds \int_0^{2\pi} d\Theta\ h(s)f(\underline{r})\ \exp\ (\mu(\underline{\Theta}'\cdot(\underline{r}-\underline{\rho}))\delta((\underline{\rho}-\underline{r})\cdot\underline{\Theta}-s) \tag{8}$$

After carrying out the integral over s we obtain

$$f_1(\underline{\rho}\ ) = \int d\underline{r}\ p(\underline{\rho}-\underline{r})\ f\ (\underline{r}) \tag{9}$$

where

$$p(\underline{\rho}-\underline{r}) \triangleq \int_0^{2\pi} d\underline{\Theta} h(\underline{\Theta}\cdot(\underline{\rho}-\underline{r}))\exp(\mu\underline{\Theta}'\cdot(\underline{r}-\underline{\rho})) \tag{10}$$

we see that f_1 is related to f by a transformation which is shift invariant with respect to the activity distribution. $p(\underline{r})$ is the point-spread function of the process, and it is radially symmetric,

$$P_1(|\underline{r}|) = P(\underline{r}) \tag{11}$$

and by evaluation we see that

$$p_1(R) = \int_0^{2\pi} d\Theta\ h(R\cos\Theta)\ \exp\ (-\mu R\sin\Theta) \tag{12}$$

To convert the above analysis to a reconstruction theorem we must find a $h(\cdot)$ for which

$$p_1(R) \cong \frac{\delta(R)}{R} \tag{13}$$

While we have been unable to obtain an analytic solution to this equation, we have been able to construct $h(\cdot)$ with sufficient numerical accuracy to obtain attractive results in the computer simulation of the process.

The above analysis suggests the following reconstruction algorithm.

a) The data is collected over a grid of p and Θ values.

$$g^*(i,j) = g\ (i\Delta p,\ j\Delta\Theta) \tag{14}$$

$$-N \leq i \leq N \qquad \Delta p = D/2N$$

$$1 \leq j \leq M \qquad \Delta\Theta = 2\pi/M$$

where D is the maximum diameter of the section being scanned. Defining a discrete distance function:

$$d^*(i,j) \triangleq d\ (i\Delta p,\ j\Delta\Theta) \tag{15}$$

b) We compute the corrected projection data.

$$g_1^*(i,j) = g^*(i,j)\exp(\mu d^*(i,j)) \tag{16}$$

c) Then convolve it with a weighting function

$$g_2^*(i,j) = \Delta p \sum_{K=-N}^{+N} g_1^*(k,j)\ h^*(i-k) \tag{17}$$

d) And finally compute the exponential backprojection

$$f_1(x,y) = \Delta\Theta \sum_j g_2(i(x,y,j),j)\exp(-\mu\ell(x,y,j)) \tag{18}$$

where

$$i(x,y,j) \triangleq (x \cos j\Delta\Theta + y \sin j\Delta\Theta)/\Delta p \tag{19}$$

and

$$\ell(x,y,j) \triangleq +x \sin j\Delta\Theta - y \cos j\Delta\Theta \tag{20}$$

As is usual in these algorithms, if $i(x,y,j)$ is not an integer $g_2(i(x,y,j),j)$ is found through linear interpolation from the values computed in (17). Equations (18) through (20) are discrete approximations to Equations (5) through (7).

SIMULATION AND ALGORITHM DESIGN

The above procedure was tested on data which simulated radionuclide scans of a head. In some of the simulations it was assumed that the attenuating region was elliptical while in other tests, a circular attenuating region was simulated. It was assumed that the linear attenuation coefficient μ was equal to .15 cm^{-1}, simulating Tc_{99}^m attenuation in soft tissue, and that the attenuation coefficient was uniform through the region. The emission profile, $g(p,\Theta)$ from Equation (3), was computed for a set of Θ values, and for a fine set of p increments (Δp = .152 cm). To simulate the action of a collimator, these data were convolved with a Gaussian function with a FWHM of 1.3 cm. and the resulting function was sampled over a grid with spacing .609 cm. In effect, we compute

$$g^*(i,j) = \int dp\ g(p,j\Delta\Theta)\exp(-\alpha(i\Delta p-p)^2) \tag{21}$$

$$\alpha = 1.6 \qquad \Delta p = .609 \text{ cm.}$$

These choices of Δp and FWHM approximate the instrumentation of the MARK-IV scanner (4).

The data computed in this way were fed into a reconstruction algorithm specified by Equations (13) through (19). The discrete distance function $d^*(i,j)$ was determined by a geometrical analysis of the attenuating field (See Figure 1). To complete the algorithm design one must specify $\Delta\Theta$, the angle increment, and $h(i)$, the convolving function. A qualitative assessment of the process is made when one chooses the activity distribution to be a

two-dimensional δ function, $\delta(\underline{r})$. The reconstruction, in this case, is the two-dimensional point-spread function, $p(\underline{r})$, defined in Equation (10). Figure 2a shows a reconstruction of this function with h(i) similar to that used by Ramachandran and Lakshminarayanan and with $\Delta\Theta=2\pi/250$. We see that near the origin the point-spread function has the desired shape, while for increasing $\underline{r}$ the function exhibits angular oscillations. The nature of these oscillation has been analyzed for the usual convolutional algorithm (7) and they arise because of the coarseness of the $\Delta\Theta$ used in sampling $g(p,\Theta)$. Figure 2b was computed with $\Delta\Theta=2\pi/500$ and we see that the oscillations are attenuated but now we note that $p(\underline{r})$ does not tend to zero as $|\underline{r}|$ increases, but it has the shape of a bowl for large values of $|\underline{r}|$. This defect is due to the exponential backprojection. A uniform activity distribution is reconstructed in Figure 2c and the defect produced by the "bowl" artifact is quite strong.

Specifications on the point spread function may be developed by examining Equation (9). $f(\underline{r})$, the activity distribution, is assumed to be zero outside the object which has diameter D. Therefore $p(\underline{r})$ should have a pulse shape at the origin, and away from the origin it should be equal to zero. Since the largest value of $\underline{\rho}$ that is to be evaluated is bounded by (D/2), the radius of the activity distribution, and for this value of $\underline{\rho}$ the magnitude of $\underline{\rho} - \underline{r}$ in the integral may be as large as D, we conclude that the point spread function must be zero outside the pulse up to a radius D or:

$$P(\underline{r}) = 0 \text{ for } \underline{r} \text{ outside the pulse and } |\underline{r}| < D \tag{22}$$

A convolving function that satisfies this requirement was designed in the following way:

1) An initial guess of the convolving function was made. Defining the Fourier transform by:

$$\hat{h}(f) = \int ds\, h(s) \exp(-2\pi i f s) \tag{23}$$

we chose

$$\hat{h}(f) = \sqrt{(\mu^2 + (2\pi f)^2)} \cdot \exp(-\beta f^2) \qquad \beta = 6.016 \tag{24}$$

The first term in this equation governs the behavior of $\hat{h}(f)$ at low spatial frequencies, and the second term is a Gaussian filter that rolls off the frequency response at high frequencies. The value of β used produces a FWHM of 1.3 cm. with an impulse input.

2) The convolving function, h*(i), was computed by taking the inverse transform of $\hat{h}(f)$ for $p = i\Delta p$, $\Delta p = .609$ cm.

3) The point-spread function $p(\underline{r})$ was computed by reconstructing an impulse with the ECA using this h*(i). A plot of this p(x,0) is shown in Figure 3a. Note that the function consists of a pulse at the origin, and of a "bowl" for large values of $|r|$.

4) The "bowl" was separated from this function by defining

$$b(x,0) = \begin{cases} p(x_o,0) & x<X_o \\ p(x,o) & x\geq X_o \end{cases} \qquad X_o = 5.48 \text{ cm.} \tag{25}$$

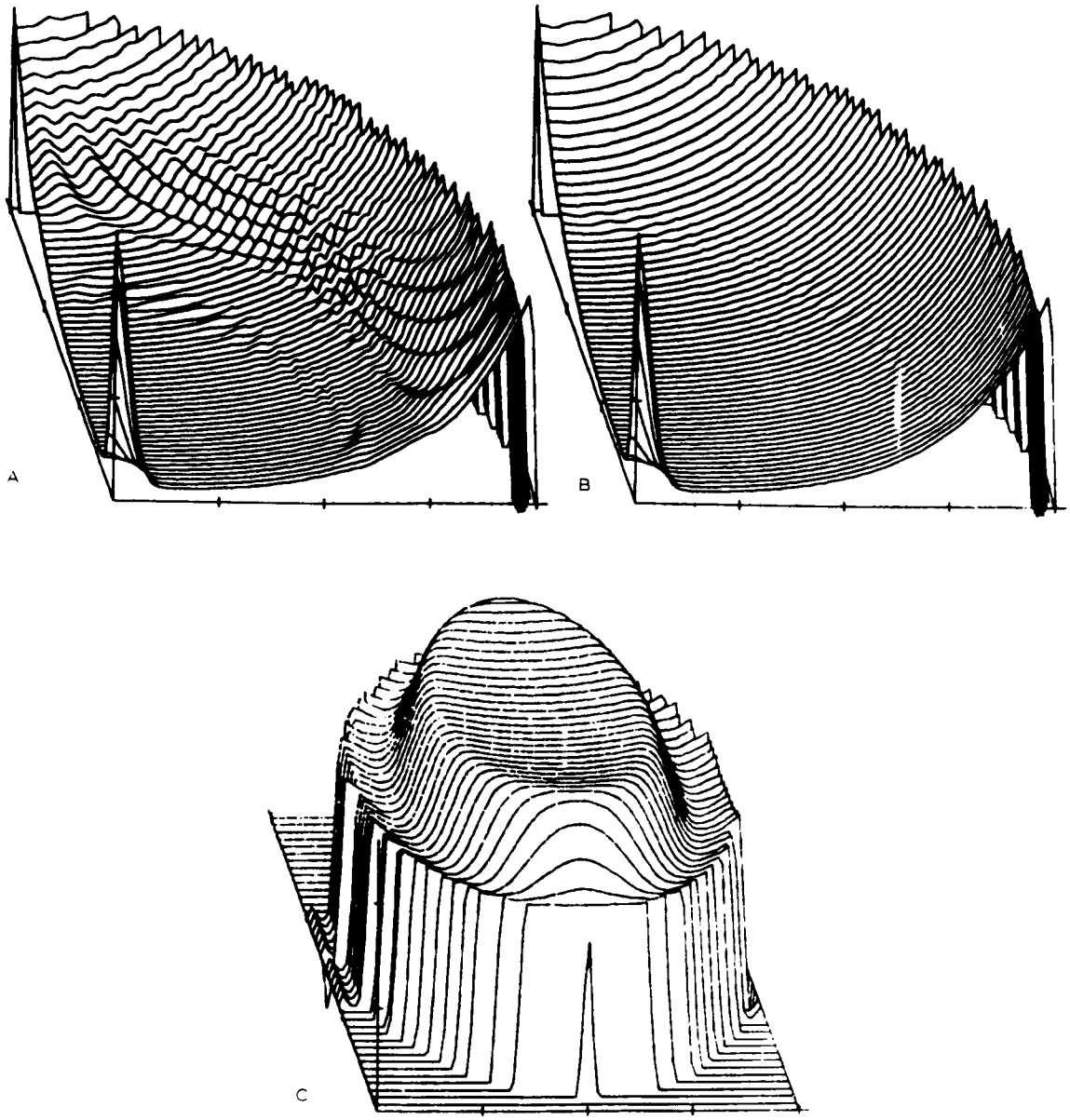


Figure 2

a) Impulse Reconstructed with Incorrect ECA. R_{max} = 38cm. Insufficient Projections Produce Angular Artifacts.

b) Same as (2), but with 500 Projections.

c) 18cm x 14cm Uniform Field Reconstructed with the Above System and 66 Projections.

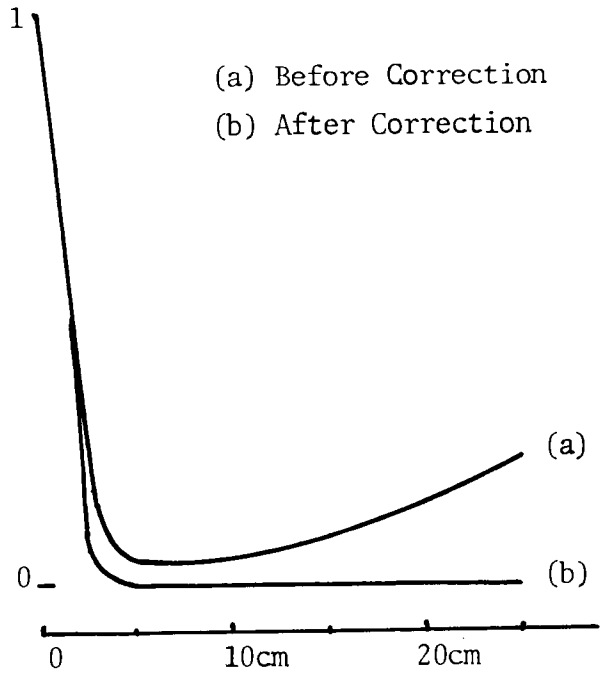


Figure 3

ECA Impulse Response

5) After several trials, it was found that the following procedure produced an appropriate correction. A uniform activity field of a circular region with a diameter of 70 cm. was assumed, and simulated projection data (diameter 76.8 cm.) was obtained with $\mu = -.15$, the negative of the attenuation coefficient used elsewhere in this study. When these projections were reconstructed with ECA, the reconstructed field matched very closely the shape of the "bowl" obtained in step 3). The data from this simulation prior to the exponential backprojection, $g_2^*(i,1)$ of Equation (17) was scaled by an appropriate constant so that this "bowl" matched that of step 3) and the scaled $g_2^*(i,1)$ was subtracted from the convolving function, $h^*(i)$, obtained in step 2). The impulse response obtained with the corrected $h^*(i)$ is shown in Figure 3b. This modified function turned out to be suitable for accurate convolutional reconstruction. It is listed in Table I. Since the function is even, it is tabulated only for positive i.

1.00763	0.195596	-0.301805	-0.135992	-0.727500E-01
-0.376417E-01	-0.318870E-01	-0.203060E-01	-0.195144E-01	-0.137021E-01
-0.137585E-01	-0.101951E-01	-0.103711E-01	-0.789926E-02	-0.803857E-02
-0.618187E-02	-0.626331E-02	-0.479752E-02	-0.483482E-02	-0.364417E-02
-0.365828E-02	-0.267779E-02	-0.268762E-02	-0.187742E-02	-0.189713E-02
-0.122968E-02	-0.126816E-02	-0.722136E-03	-0.783491E-03	-0.340655E-03
-0.425050E-03	-0.688555E-04	-0.173613E-03	0.111174E-03	-0.943032E-05
0.217944E-03	0.867838E-04	0.269220E-03	0.132890E-03	0.281186E-03
0.144530E-03	0.267643E-03	0.134593E-03	0.239654E-03	0.113083E-03
0.205481E-03	0.871996E-04	0.170799E-03	0.616704E-04	0.139056E-03

Table I

The First 50 Values of the ECA Weighting Function

An activity field that simulates a head in Figure 4a. The outer source field, modeling the blood supply surrounding the brain, was elliptical with

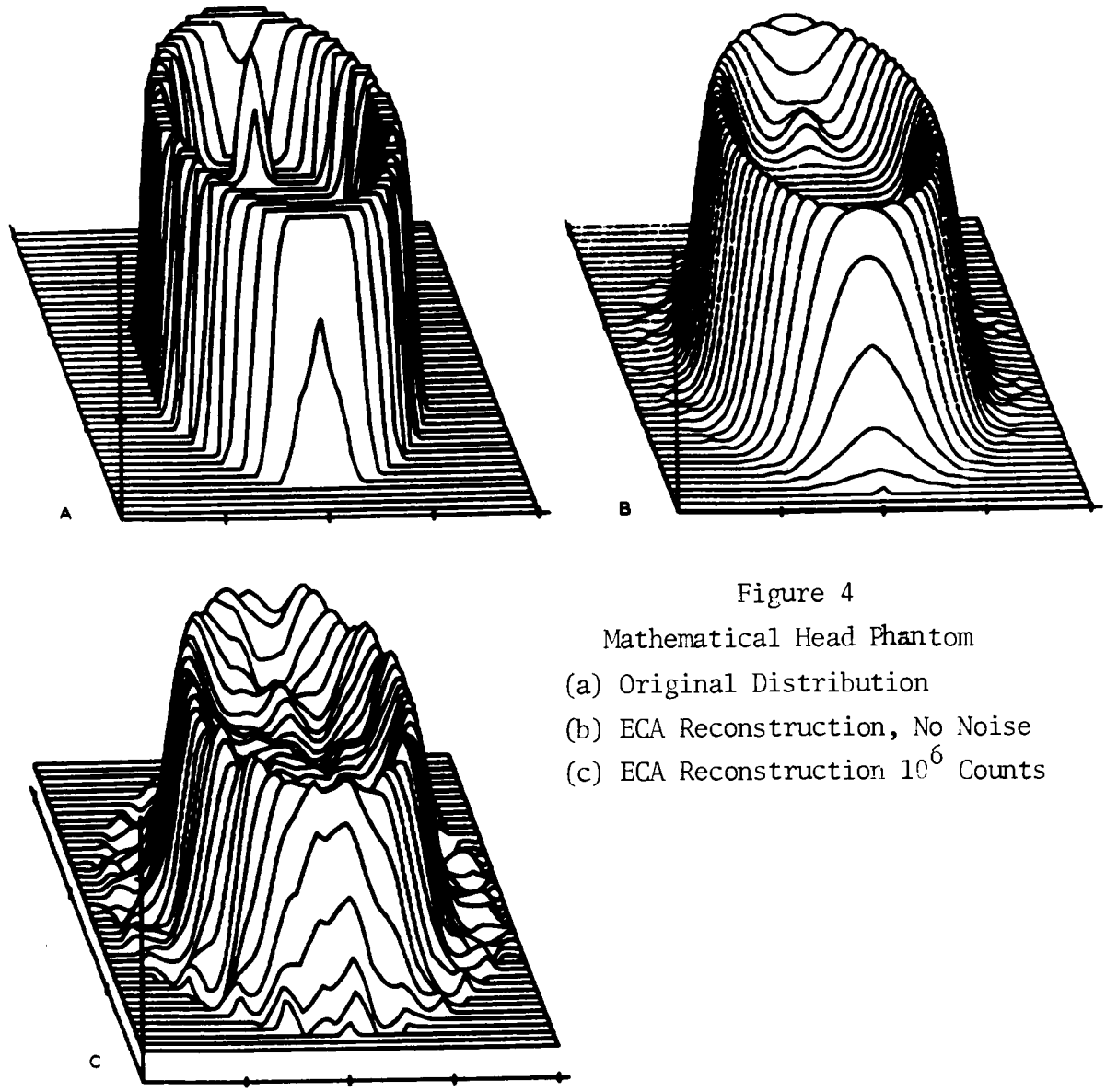


Figure 4

Mathematical Head Phantom

(a) Original Distribution

(b) ECA Reconstruction, No Noise

(c) ECA Reconstruction 10^6 Counts

major radius 9.779 cm., minor radius 7.67 cm., and a relative activity of 1. An interior source field, modeling the brain, was elliptical with radii 7.779 and 5.67 cm., and a relative activity of -1/2.This was added to the outer source field to yield an activity 1/2 that of the blood supply. The tumor centered at (0, 5.49) had a radius of .65 cm., and increased the relative activity of this part of the brain to that of the surrounding blood supply. The reconstruction of this section is shown in Figure 4b. This was computed with $\Delta\Theta = 2\pi/66$.

Noise Considerations

Because the counting statistics in ECAT are fairly poor, noise requires serious consideration. The following procedure was used to account for noise in our simulations.

Let $g^*(i,j)$ be the simulated scan data. We assume that the total counts collected in the study is C_0, and therefore the number of counts in the(i,j)'th measurement is

$$c(i,j) = \frac{C_o \, g^*(i,j)}{\sum_k \sum_\ell g^*(k,\ell)} \tag{26}$$

Assuming Poisson statistics, the standard deviation of $c(i,j)$ is $\sqrt{c(i,j)}$, so that the standard deviation of $g^*(i,j)$ due to the noise is

$$\sigma(i,j) = \sqrt{\frac{g^*(i,j)\sum_k\sum_\ell g^*(k,\ell)}{C_o}} \tag{27}$$

The noise in $g^*(i,j)$ is simulated by generating the noise set

$$n(i,j) = \sigma(i,j)r(i,j) \tag{28}$$

where $r(i,j)$ is a set of independent Gaussian pseudorandom numbers with zero mean and unit standard deviation. A simulated noisy data set may be obtained by forming

$$g_n^*(i,j) = g(i,j) + n(i,j) \tag{29}$$

and a simulated noisy reconstruction may be obtained by entering $g_n^*(i,j)$ into the reconstruction algorithm. Since the algorithm is linear, the same reconstruction can be obtained by entering $n(i,j)$ into the reconstruction algorithm and adding the reconstructed values, $n_o(x,y)$, to the noise-free reconstruction $f_1(x,y)$ so that the noisy reconstruction is

$$f_n(x,y) = f_1(x,y) + n_o(x,y) \tag{30}$$

The signal to noise ratio of the reconstruction is defined as

$$(S/N)\ (x,y) = \frac{f_1(x,y)}{\sigma_o(x,y)} \tag{31}$$

with $\sigma_o(x,y)$ being the standard deviation of $n_o(x,y)$ which may be computed by reconstructing a number of noise data sets generated with different seeds in the pseudorandom number generator and estimating the standard deviation of the reconstruction at coordinates (x,y).

The (S/N) ratio depends on the object being reconstructed, the (x,y) coordinates, the convolving function, and on the total counts observed. To eliminate the dependence on the number of counts, we define the reference S/N ratio of a particular object and algorithm as the signal to noise ratio with $C_o = 10^6$

$$(S/N)_r\ (x,y) \triangleq S/N\ (x,y)\ \Big|_{C_o = 10^6} \tag{32}$$

For the same object, the S/N ratio with total counts equal to C is given by

$$(S/N)\ (x,y) = (S/N)_r(x,y)\sqrt{\frac{C}{10^6}} \tag{33}$$

An example of a simulated noisy reconstruction of the field of Figure 4a is shown in Figure 4c.

EVALUATION

The ECA algorithm was evaluated by comparing it to two other algorithms. This evaluation was carried out with simulated data. The two other algorithms that were used in this comparison are based on the convolutional algorithm,

but they contain ad-hoc methods for attenuation correction. One of these is the pre-correction algorithm with is based on a concept published by Kuhl, et al. (4). Let the scan data be represented by the array $g^*(i,j)$. An even number of projection angles is used and data from opposing views is grouped together so that

$$g^*_a(i,j) = g^*(i,j)+g^*(-i,j+\frac{M}{2}) \qquad (34)$$
$$j = 1 \ldots \frac{M}{2}$$

These grouped data are corrected by a set of multiplicative weighting factors

$$g_b(i,j) = g^*_a(i,j)\ w_1(i,j) \qquad (35)$$

and the weighting factors are chosen as follows: Let $g_o(i,j)$ be the scan data from a uniform distribution, computed for attenuation coefficient μ equal to zero, and let $g_\mu(i,j)$ be the scan data obtained from the same uniform distribution in the presence of attenuation. The correction coefficients are

$$w_i(i,j) = g_o(i,j)/g_\mu(i,j). \qquad (36)$$

In the evaluation, an activity distribution which was uniform over a disc of diameter 18.812 cm. was used. The attenuation field was circular and had a diameter 1 cm. larger. The corrected data $g_b(i,j)$ is input to a convolutional reconstruction algorithm producing a reconstruction $f_2(x,y)$. The rationale for this procedure lies in the fact that if $g^*(i,j)$ is obtained from the reference uniform object, then the reconstruction will be as faithful as the capability of the convolutional algorithm allows.

The second procedure used for comparison is the post-correction algorithm based on the method described earlier by Kuhl, et al. (1). The projection data from opposing views is grouped as in the pre-correction algorithm, and the grouped data are reconstructed with the convolutional algorithm producing an uncorrected reconstruction $f_\mu(x,y)$. A multiplicative correction is applied to this producing a reconstruction,

$$f_2(x,y) = f_\mu(x,y)\ w_2(x,y) \qquad (37)$$

and $w_2(x,y)$ is computed as the ratio of $f_o(x,y)$, the convolutional reconstruction of the attenuation-free data $g_o(i,j)$ and of $f_\mu(x,y)$ which is the convolutional reconstruction of $g_\mu(i,j)$.

$$w_2(x,y) = \frac{f_o(x,y)}{f_\mu(x,y)} \qquad (38)$$

This algorithm will also reconstruct without error the data from a uniform distribution of the same dimensions as that used for determining the correction factor.

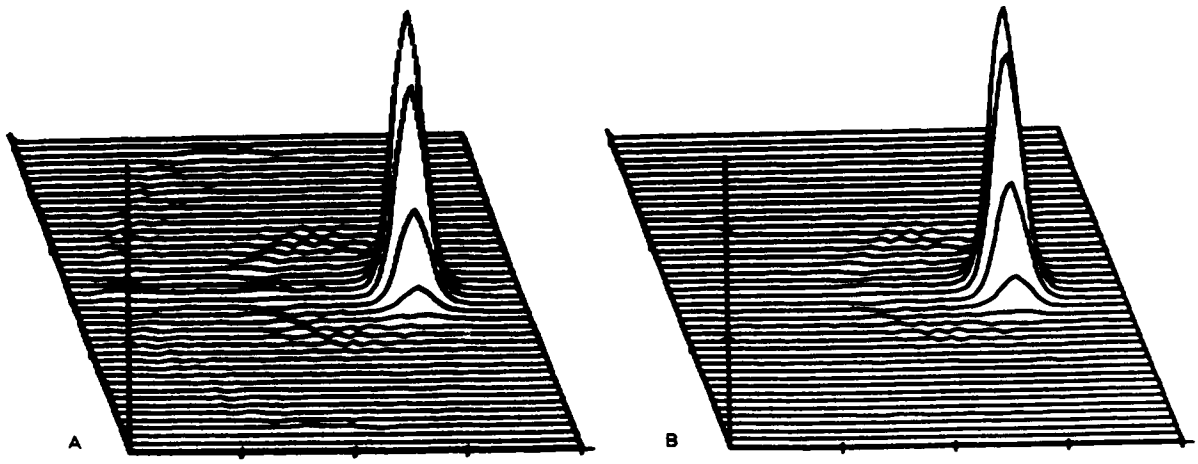
In both the pre- and post-correction algorithm the convolutional weighting function was chosen to have the Fourier Transform

$$\hat{h}(f) = |f|\exp(-\alpha f^2) \qquad (39)$$
$$\alpha = 6.016$$

so that the resolution for these algorithms was approximately equal to that of the ECA.

The comparison between the algorithms was made along two aspects: the fidelity of the reconstruction, and the S/N ratio Fidelity was measured by

reconstructing point sources of activity at various positions in the field. Figure 5 shows reconstructions of impulses located 6.1 cm. from the center. In qualitative terms, all three algorithms reconstruct the point radiator as a nearly Gaussian pulse.



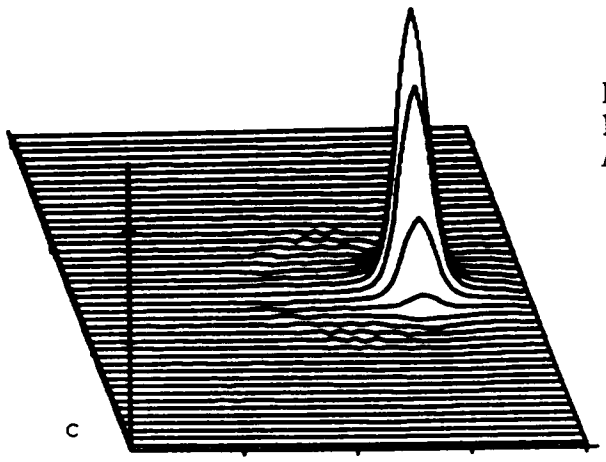


Figure 5

Reconstructions of an Impulse 6cm From Center in a 19.9 Diameter Attenuating Field.

(a) ECA

(b) Pre-Correction

(c) Post-Correction

A quantitative comparison is made by computing the total area under the reconstructions produced by the three algorithms and plotting this area as a function of distance between the impulse and the center of the field. These results are shown in Figure 6. We see that both the pre- and post-correction

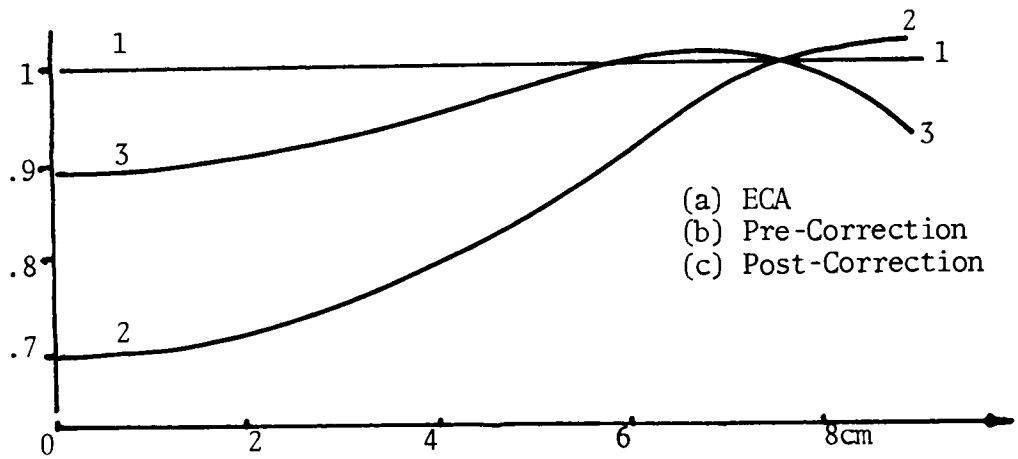


Figure 6
Point Source Uniformity in a 20cm Diameter Cylinder

algorithms depart from perfect fidelity, and the departure depends on the position of the impulse. These calculations were made for a circular attenuating field whose diameter was 20 cm.

The noise gain evaluation was carried out with a round test field of uniform activity with diameter 18.8 cm. embedded in a uniform attenuation section of diameter 19.8 cm. The noise gain factor was computed with an analytic method such as that used in the analysis carried out by Shepp and Logan (8). The S/N ratio, as a function of radius, is plotted in Figure 7. All three

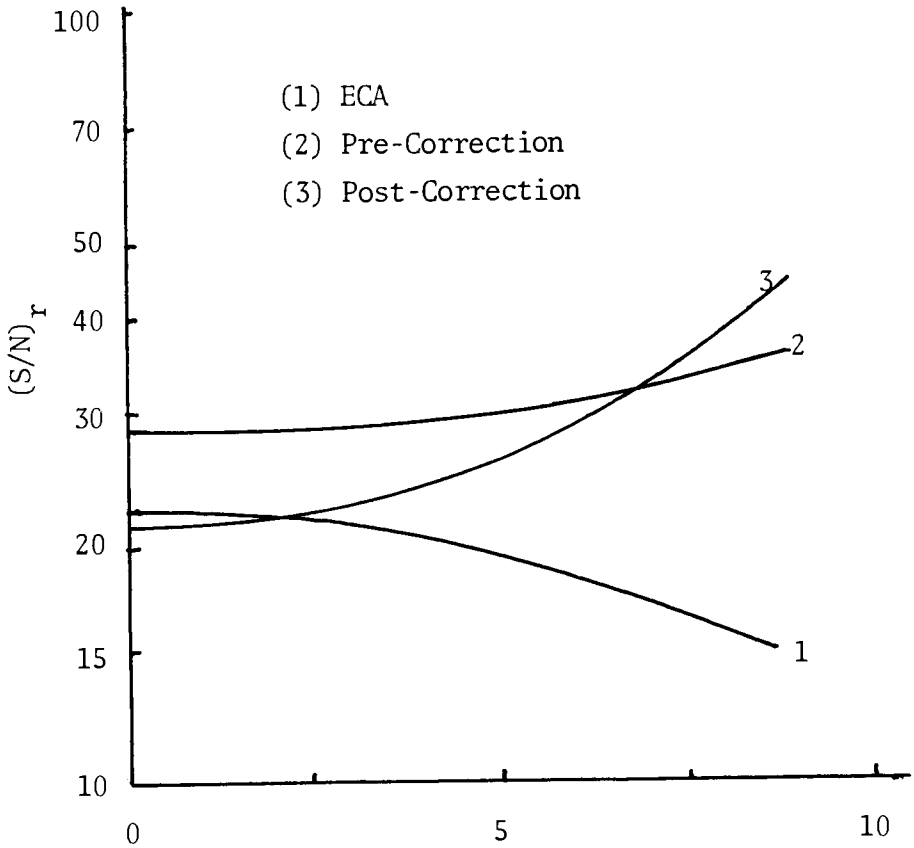


Figure 7

Signal-to-Noise Ratio for a Uniform Distribution in an 18.8 cm Diameter Cylinder with 10^6 counts.

algorithms show similar S/N ratio at the center, but as radius increases the behavior of the three algorithms is quite different. The S/N ratios for both the pre- and post-correction algorithm increase, while that for the ECA decreases.

DISCUSSION

The Exponential Backprojection Algorithm is based on an analytic inversion result and it has been shown to correct distortion due to attenuation in emission computed axial tomography. The evaluation was carried out on computer simulated scan data.

ECA has some shortcomings. Perhaps the most serious is the inferior S/N ratio near the edges of the field. A further limitation is the fact that the attenuation field must be uniform. Of considerable interest is a possible ad-hoc algorithm in which the backprojection factor is based on an actual,

nonuniform, attenuation factor distribution. Perhaps the greatest surprise in this study was how well the post-correction algorithm performs. The current data suggests that the ECA is most suitable for application in which the attenuation field is uniform and where utmost accuracy is required.

REFERENCES

1. Kuhl, D.E., Edwards, R.Q., Ricci, A.R., and M. Reivich, "Quantitative Section Scanning Using Orthogonal Tangent Correction", J. Nucl. Med., 14: 196-200 (1973).

2. Budinger, T.F., and Gullbert, G.T., "Three-Dimensional Reconstruction in Nuclear Medicine Emission Imaging", IEEE Trans. NS21:2-20, (1974).

3. Kay, D.B., and Keyes, J.W., Jr., "First Order Corrections for Absorption and Resolution Compensation in Radionuclide Fourier Tomography", J. Nucl. Med. 16: 540-541, (1975).

4. Kuhl, D.E., Edwards, R.Q., Ricci, A.R., Yacob, R.J., Mich, T.J., and Ahlavi, A., "The Mark IV System for Radionuclide Computed Tomography of the Brain", Radiology, 121:405-413, (1976).

5. Walters, T., Simon, W., Chesler, D.A., Correia, J.A., and S.J. Riederer, "Iterative Convolution for Radionuclide Axial Tomography Corrected for Internal Absorption", Proc. IVth Int. Conf. on Inf. Proc. in Scintigraphy pp. 333-342, Orsay, France (1975).

6. Ramachandran, G.N., and Lakshminarayanan, A.V., "Three-Dimensional Reconstruction from Radiographs and Electron Micrographs: Part III-Description and Application of the Convolution Method", Ind. J. of Pure and Appl. Sci. 9:997-1003, (1971).

7. Tretiak, O.J., "The Point-Spread Function for the Convolutional Algorithm" Image Processing for 2-D and 3-D Reconstruction from Projections, Technical Digest, Optical Society of America, August 1975, Th AS.

8. Shepp, L.A., and Logan, B.F., "The Fourier Reconstruction of a Head Section", IEEE Trans., NS21: 21-43.

DISCUSSION:

Oppenheim commented that the exponential weighting carried out in the back-projection step tremendously increases the noise for deep objects. He pointed out that in reconstruction from singles it is convenient to average opposing projections, which counteracts the attenuation effect for deep objects. He asked whether this approach could be incorporated into ECA. Tretiak said that exponential back-projection was used in order to obtain shift invariance. He speculated that there should be some optimal algorithm which also gives shift invariance but produces noise throughout the reconstruction which is the same as the noise at the center.

Muehllehner noted that attenuation was assumed constant, and asked what would happen when attenuation varies. Tretiak responded that the problem has not been investigated yet and felt that the method should still work, although not as faithfully.

IAEA Image Processing Intercomparison

Chairman: Stephen M. Pizer (Chapel Hill, NC, USA)

Andrew Todd-Pokropek and Charles Metz presented the objectives, procedures, and results of the IAEA Intercomparison, which was designed to determine which scintigraphic image processing techniques were preferable for the detection and localization of small lesions and whether any of these techniques produced improvement over unprocessed analog images. The complete report can be found on pp. 585-615 of Volume 1 of Medical Radionuclide Imaging, IAEA, Vienna, 1977 in an article entitled "IAEA Co-ordinated Research Programme on the Intercomparison of Computer-Assisted Scintigraphic Techniques, Third Progress Report". Andrew Todd-Pokropek indicated that besides him and Charles Metz, credit for the intercomparison should go to Eberhard Jahns, Lars Johansson, and Hugh Belcher. The following summarizes the presentation.

After having completed an intercomparison using computer-simulated images of a circularly symmetric object, it was decided that a study should be carried out with images of a real phantom with considerable normal structure, since simulated images were not realistic enough and in clinical images, even with simulated lesions, the correct answers were not fully known. The phantom used was a mold cast from a human cadaver liver and filled with pertechnetate solution. Cold spots and hot spots were created by suspending, in the phantom, spheres which were voids or held more concentrated pertechnetate solution. Using a gamma camera with a high resolution collimator, analog images of the phantom with lesions were recorded on x-ray film using a small-dot multiformatter, and this same data was digitized into a 128 x 128 matrix (a few of the participants compressed this to 64 x 64) and stored on magnetic tape. Forty-six phantom configurations with 0-6 hot spots (with a mode at 3) and fifty phantom configurations with 0-6 cold spots (with a mode at 3) were imaged, each with the spots in random positions. The lesions were of five sizes between 1.2 and 3.5 cm in diameter with a mode at approximately 2.0 cm. The positions of the lesions were precisely measured using x-rays of the phantom. In some of the configurations the phantom was tilted to simulate normal biological variation and to attempt to defeat processing techniques based on producing a template of the "normal" phantom. The ninety-six images produced simultaneously in both analog and digital form had a maximum count density of approximately 1600 counts/cm^2. Images with a maximum count density of 2400 counts/cm^2 were also made of the 96 configurations, but only in analog form. The purpose of these images was to allow measuring the improvement obtained by increased count density in the analog images in order to calibrate the changes due to processing which were to be measured.

The analog images on x-ray film were sent to institutions experienced in reading analog images, and magnetic tapes with the digital data were sent to different institutions experienced in processing scintigrams and reading images from digital displays. Participants receiving the digital images processed the images by their usual techniques and displayed them on their usual displays, and many separately read the series of images on their digital displays with unprocessed data. All of the participants were told the point spread function of the gamma camera and the distribution of number of lesions per image. They were also told for each image whether it contained hot spots or cold spots. They were instructed to train themselves by looking at all the images before reading any, to determine by constant features the normal structure, which was often more obvious than the lesions. They were to report the position of each suspected lesion and a confidence rating for each. Using ROC techniques, results were analyzed from 9 observers of the analog images with 1600 counts/cm² maximum (low count density), 4 observers of the analog images with 2400 counts/cm² maximum (high count density), and 56 observer x processing technique combinations from 13 institutions with the digital images. Details of the analysis techniques will not be given here, but it is worth noting that some new developments in ROC techniques were made for this analysis.

The results, in brief summary, were that the analog high count density images were better than the analog low count density images, which were about equally as good as the digital (low count density) images processed using mid-frequency enhancement filters, which were better than the digital images processed using either low pass (smoothing) or high pass (edge enhancement) filters, which were better than the unprocessed digital images. Thus, with the digital displays used, none of which are state of the art today, processing did not improve images over the unprocessed analog images, but especially the mid-frequency enhancement filters improved them greatly over the unprocessed digital images (this last agreeing with results from the earlier study with simulated images with a simple normal structure). It appears that the display has more effect than the processing and that further work needs to be done to intercompare displays and to understand the relationship between displays and processing techniques, as different displays appear to be optimal for different processing techniques.

In discussion of the procedures used in the study, the following points were made. Michael Goris questioned whether the Picker camera used was a representative device, and Andrew Todd-Pokropek said he thought is was. To Alan Ashare's question as to how the observers distributed their responses among the five confidence categories, Charles Metz answered that though the instructions were to calibrate confidence so as to use the categories about equally, some observers did not. It was pointed out that it was possible for observers to calibrate their confidences because in ROC rating experiments the absolute confidences do not matter, only that they are used consistently. Brian Pullan suggested that if observers had been encouraged to use more than five categories of confidence, equalization into five categories could have been done aft - the fact as part of the analysis.

Ivan Runczik noted that not enough normal variation had been imposed; template matching worked too well. He and P.R. Bell wondered whether the large "normal" variations weren't greater, compared to lesions, than in clinical studies, but the speakers said they felt that they were as great in clinical studies. Brian Pullan said the experiment was poorly designed because there were too many variables to control with such a complicated phantom, but Andrew Todd-Pokropek answered that different results are obtained with simpler phantoms, and Stephen Pizer pointed out that results with a simple phantom would not be convincingly extendible to clinical images.

The discussion went on to the results of the study. Michael Goris asked why the range of ROC curves for the analog images was smaller than for digital images even for similar (e.g. no) processing, and Charles Metz responded that the consistency of the display with the analog images and large variation in displays with digital images would explain that fact. Lloyd Knowles emphasized the result with analog images that significant improvements are obtainable by increasing the maximum count density from 1600 to 2400 counts/cm^2, noting that the old standard of 800 counts/cm^2 was clearly too low.

The remaining discussion centered on the issue of the importance of displays in determining image quality. Brian Pullan described an ROC experiment carried out by his group in Manchester comparing the viewing of CT scans 1) on polaroid film, 2) immediately in front of the display station using its interactive features, and 3) sitting 20 ft. from the display station such that the solid angle subtended by the display at the eye was the same as with polaroid images and no interaction was possible. Distant use of the display station gave the best results, and interactive use of the display station gave the worst! He stated that both the viewing angle and the capability of interaction had an effect on this result. He also pointed out how important an effect this study showed the choice of display to have.

After Ivan Runczik commented that the IAEA study showed that either image processing should be abandoned or the weak link identified, Stephen Pizer noted that the strong degradation from unprocessed analog images to unprocessed digital images showed the importance of improved digital displays. He said that it was not clear whether, if one had a digital display as good as the analog display, the strong improvement shown to be the case for processed digital images over unprocessed digital images would continue to hold or whether processing was simply compensating for the digital display. His avowedly poorly based feeling was that with improved displays, processing would still be shown of value, i.e. to produce significant improvements over analog images. Many agreed that work on displays was required, though it was pointed out by Charles Metz that this would involve findng optimum displays for each of different processing techniques (including no processing).

A number of discussants suggested that the processed IAEA images from the various centers should be returned to a central location for comparison on a single display and by a fixed set of observers, so as to remove the uncontrolled effects of observer ability and display quality on the results of the IAEA study.

However, many saw that the problems with this idea were that it was not obvious how to choose the fixed display (a type of processing, as noted by Goris) and that the results might be biased towards processing techniques for which the display was especially well matched. Keith Britton and Leslie Rogers suggested that all digital images should be on film like the analog images, but it was pointed out that the problem remained of what to put on the film.

It was agreed that what was needed was for the processed IAEA images to be sent in digital form to a central organization which would distribute them to observers with different display systems, who would read the displayed results and send the readings to the central organization for scoring and analysis. By this method the effectiveness of various combinations of processing technique and display could be understood and standards could be developed. It was pointed out that sharing the IAEA images to allow further studies had begun in an informal way, e.g. Andrew Todd-Pokropek was using the unprocessed images to compare gray-scale and color displays, but the processed images had not yet been collected. A protocol for collection, distribution, and observer data scoring and analysis had been defined but not yet implemented. It had been agreed in principle by the IAEA that BCTIC would serve as the central organization to serve as clearinghouse and score observer results, but the IAEA had not yet agreed to continue the analysis of this data. The strong consensus was that such analysis was an important task.

COMPUTER DISPLAY OF ORGAN SURFACES[1]

Gabor T. Herman[2] and Hsun Kao Liu[3]
State University of New York at Buffalo
Buffalo, New York, U.S.A.

ABSTRACT

Methods of hidden surface removal and shading for computer displayed surfaces are proposed. If the surface to be displayed is approximated by a large number of square faces of restricted orientation, the methods proposed in this note work at least an order of magnitude faster than previously published methods. A situation where such approximation is natural is the display of organs based on data obtained by the reconstruction of the internal structure of a body from multiple shadowgraphs such as x-ray photographs (computerized tomography). The methods can also be used for rapidly achieving some but not all of the aims of surface modelling in more general situations. No sorting is required for the hidden surface removal and the artifact caused by approximating curved surfaces with polygons is reduced by low-pass filtering of the picture to be displayed. The results are demonstrated by displays of actual organs based on data provided by computerized tomography.

INTRODUCTION

Our aim is to display the precise three-dimensional appearance of organs which are inside a body (e.g., the heart or the lungs) without having to open up the body. This is achieved by a sequence of the following steps.

[1]The research for this work is supported by NSF grant MCS75-22347, NIH grants HL-18968 and HL-4664 and NCI contract CB-53860.

[2]Department of Computer Science, State University of New York at Buffalo, Amherst, NY 14226.

[3]Department of Mathematics, California State Polytechnic University, Pomona, CA 91768.

(i) X-rays are taken of the body from a number of directions, and from the data a three-dimensional array of numbers is produced, representing the internal density distribution of the body. This process is called reconstruction from projections [5,6,10]. Essentially the same mathematical procedures can be used in reconstructing the internal structure of molecules from electron-micrographs, the solar atmosphere from coronameter data, as well as many other areas in medicine, science and engineering (ultrasound, nuclear medicine, non-destructive testing, holographic interferometry, radio astronomy, etc., see the references in [5]). In any case, a three dimensional array of numbers is produced, in which each number is an estimate of the average of some physical parameter (e.g., density) in one of a set of equal, non-overlapping cubes, called voxels (volume elements). The nature of the voxels is important for what follows. In particular, they must be small compared with the features we desire to resolve.

(ii) A particular organ is selected (by specifying a voxel near its boundary), and the surface of the organ is detected using a three-dimensional boundary detection algorithm [12]. (In case a time sequence of reconstructions is available, a four-dimensional boundary detection algorithm, which makes essential use of the continuity of changes over time, can be used to detect the time sequence of a changing surface [11].) Naturally, such a surface detection algorithm will in the first instance produce a surface which follows the faces of the voxels. Such a surface consists of a large number of elements of equal size, each one of which is a square parallel to one of three mutually perpendicular planes. Bearing in mind that the detected surface must be a close approximation of the true surface of the organ, so that medically significant information is not lost, we find that the number of square-shaped surface elements needed to describe the surface of an organ such as the heart or the lung is typically 10,000 - 15,000. The detected surface could now be approximated by larger polygons other than squares (or even more complex surface elements such as bicubic patches), but since we are interested in accurate quantitative evaluation of a particular organ (rather than just what a heart or a lung looks like in general), the resulting loss of accuracy may not be justified. (In particular, the use of voxel faces as surface elements makes volume measurements, often used in quantitative diagnostic medicine, trivial.) We shall look again at the relative advantages of using a large number of simple or a small number of more complex surface elements in our discussion section below. An alternative approach to truly three-dimensional surface detection of organs, such as the one described in [12], is detection of two-dimensional boundaries in cross sections followed by the connection these boundaries by planar surface elements such as triangles [8,9,13]. This has several drawbacks: information provided by the continuity of the organ surface is lost, bias is introduced into the display by the direction of the original cross sections, concavities in the organ are often missed, and the display of the resulting surfaces requires a considerable amount of computer processing.

(iii) The detected surface of the organ is displayed. How best to do this under the given circumstances is the subject matter of this note.

METHODS

In order to achieve a visually realistic representation of the surface of a three-dimensional object on a cathode-ray tube, one has to solve the problems of removing the hidden parts of the surface and shading its visible parts. The approximation of the surface so that it is composed of planar polygons is a commonly used approach which allows efficient solutions to both these problems [3,7,8,9,13,14,15,16]. Lack of continuity in the shading across polygonal boundaries causes undesirable artifacts which can be removed by continuous shading procedures [7,14,15].

The methods that we discuss in this section make essential use of the special properties of our surface elements: they are all squares of the same size, lie in one of three mutually perpendicular directions and are small. Previously proposed algorithms for hidden surface removal and shading apply to surfaces made up from more complicated elements. Making use of the special properties of our surface elements, we can achieve a three-dimensional display of the organ which is as good or better in quality than those produced by previously proposed algorithms; and we can do this in a time which is at least an order of magnitude faster than what would be needed by the more general algorithms if they were applied to the same surface.

In order to make precise the nature of our surface elements, we introduce the notion of a "cuberille". This is a special case of the more general notion of a "dissection" of three-dimensional space, which is defined and discussed in the Appendix.

<u>Definition</u>. A <u>cuberille</u> S is a pair $\{\Phi, C\}$, where
(i) Φ is an infinite collection of closed cubes (called the <u>components</u> of Φ) which cover the whole of three-dimensional space and which are such that if two components of Φ intersect then the intersection is a face of both components, and
(ii) C is a mapping which associates with every face the point at the center of the face.

The motivation for this name is the following. "Quadrille" marking on paper is obtained by two orthogonal sets of equally spaced parallel lines, which dissect the surface of the paper into equal squares. Analogously, a cuberille is a dissection of space into equal cubes by three orthogonal sets of equally spaced parallel planes. Note that what we call a "face" in the cuberille is always a face, in the usual sense, of exactly two cubes. The domain of the mapping C is the set of all faces, its range is the set of all points which are at the centers of the faces of the cubes. In what follows we often talk about the "plane of a face α"; by this we mean the infinite plane in which the face α lies.

In this terminology, the voxels are components of a cuberille, S, and all the surface elements are faces of S. A property of cuberilles, which will be of significant help in hidden surface removal, will be stated and proved below.

All illustrations in this work have been produced without the use of perspective transformations. Since the objects we are working with (animal organs) are usually small relative to their dis-

tance from the simulated observer, perspective transformations do not appreciably improve the quality of our displays. On the other hand, quantitative information is lost, since distances between parts of the organ are distorted. For this reason, we have abandoned using perspective transforms, which has also resulted in reducing computing costs.

One important departure from commonly used improvements to the visual appearance of the display [7,14,15] is the way we handle smoothing across the polygonal boundaries. In essence, we delay the smoothing until everything else has been done, and we smooth the final image by using a low-pass filter. We shall return to this point later on.

Right until the final smoothing, an individual visible surface element is shaded in the same way as it would be if it appeared in isolation. The exact shading rule does not appreciably affect the efficiency of our methods. (In the examples below we used a variant of Warnock's shading rule as given on p. 624 of [7], in which shading depends on the cosine of the angle the face makes with an assumed direction of light and on the distance of the point from an assumed light source). What is important is that (i) due to the smallness of the surface elements we may assume that the shading assigned to any point on the face α is the same as the shading assigned to its center $C(\alpha)$ (thus shading needs to be calculated only once for each face), and (ii) the faces lie only in three possible directions, making the calculation of the shading very fast if light rays are assumed to be parallel.

The next important consequence of our assumptions is a great help in reducing time for hidden surface removal. This consequence is stated in the following theorem, the proof of which is given in the Appendix. In the statement of the theorem $|PQ|$ denotes the distance between the points P and Q in three-dimensional space.

<u>Theorem</u>. If

(i) $S = \{\Phi, C\}$ is a cuberille,
(ii) O is a point,
(iii) α and β are two different faces of S, such that O does not lie in the plane of either α or β,
(iv) A is a point in the interior of α,
(v) B is a point in the interior of β,
(vi) O, A and B are collinear with A and B on the same side of O,

then

(vii) $|OA| < |OB|$ if, and only if, $|OC(\alpha)| < |OC(\beta)|$.

The practical consequence of this theorem is the following. Suppose we have two surface elements α and β, such that O does not lie in the plane of either α or β and the point A in the interior of α lies on the same line of sight from the observer O as the point B in the interior of β. It follows from our theorem that A hides B if, and only if, the center $C(\alpha)$ of the face α is nearer to the observer O than the center $C(\beta)$ of the face β. (It is easy to show that such a theorem is not true for more complicated surface elements; it already breaks down if the surface elements are rectangles rather than squares.)

As a result of this theorem and the discussion preceding it, we see that, for a fixed observer position O, each surface element

α need only to have two numbers associated with it for the purpose of shading and hidden surface removal:

(i) $b(\alpha)$ - the brightness of all points on the displayed surface α, calculated by the chosen shading rule, and
(ii) $d(\alpha)$ - the distance of $C(\alpha)$ from the observer O.

The unsmoothed visible surface is calculated by the following simplified version of what is generally referred to as the Z-buffer algorithm [1,2,4].

With each dot (raster element) D on the display screen we associate two numbers: brightness b(D) and distance from the observer d(D). Initially, for each dot we assign zero (dark) to the brightness, and a very large number (essentially infinity) to the distance from the observer. If a face is such that its plane contains O, then it is not displayed. All other faces making up our surface are displayed one by one, according to the following procedure.

For each face α, we calculate $b(\alpha)$ and $d(\alpha)$, and locate all the dots on the display screen that would be used if that face alone were to be displayed. For each such dot D, we change b(D) and c(D) only if $d(\alpha)$ is less than d(D). In that case, we set b(D) to $b(\alpha)$ and d(D) to $d(\alpha)$. In this fashion, by the time we have dealt with all the dots in all the faces, brightnesses will be displayed for the visible faces, and the hidden faces will not influence the display.

The main advantage of Z-buffer type algorithms over others is its speed, mainly because it does not require any sorting of faces according to their distance from the observer. This efficiency in time is further improved in our case by the fact that the values $b(\alpha)$ and $d(\alpha)$ need be calculated only once for each of the faces. Catmull [2], pointed out two possible disadvantages of such algorithms. The first is the storage requirements for the Z-buffer, which holds the values of d(D) for each display dot D. It should be noted that, in our case, the entries in the Z-buffer will be equal to $d(\alpha)$ for some face α of the cuberille, and so only a few digits of accuracy is required to make the decision needed in the hidden surface removal algorithm. The other possible disadvantage mentioned by Catmull [2] is the problem of "anti-aliasing", a problem to which we shall return in our discussion section below.

Approximation of the curved surface with faces of a cuberille generates an undesirable visual effect, caused mainly by having only three directions in which the surface elements lie. (We show an example below. Similar, though usually less disturbing, artifacts are observable whenever a curved surface is displayed as a collection of planar polygons.) We now discuss our smoothing method which substantially reduces this artifact using very little computer time. The method is a variant of "low-pass filtering". Low-pass filters are commonly used to remove noise and sharp boundaries in images.

Let $G_1, \ldots, G_9$ denote the shading intensities in a screen dot and its eight neighbors before smoothing, as indicated by the diagram below.

G_6	G_2	G_7
G_3	G_1	G_4
G_8	G_5	G_9

After smoothing, the new intensity of the center dot will be

$$G_1^{new} = \frac{G_1 + W \sum_{i=2}^{5} f_i G_i + W^2 \sum_{i=6}^{9} f_i G_i}{1 + W \sum_{i=2}^{5} f_i + W^2 \sum_{i=6}^{9} f_i}$$

If the dot labelled by 1 is on the edge of the screen and therefore does not have a neighbor labelled by i, then f_i is assumed to be zero, otherwise f_i equals one. W is a user adjustable weighting factor.

For such smoothing to be effective, it is essential that the individual faces should occupy no more than about ten raster elements on the screen. As will be discussed below, both this problem and the problem of aliasing (due to some surface elements not covering any display dots) can be handled by using a larger smoothing array. First we give some examples produced by our method.

EXAMPLES, TIMINGS, COMPARISONS

For a demonstration of our ideas we use a plastic cast of an isolated canine left ventricle with some beads inserted on the surface. Figure 1 shows a television image of this cast, blurred so as to make the resolution similar to that obtained in the reconstruction process.

A variant of the Algebraic Reconstruction Technique ART3 [10], has produced from 34 x-ray projections of our cast a 64 x 64 x 28 density array, with the edge of each voxel being just over 1 mm. (Based on such a reconstruction, variations in the surface smaller than a voxel could not possibly be displayed. For a fair evaluation of the efficacy of our display procedure its output should be compared to an image of the original which has been blurred as in Figure 1 to remove features smaller than the resolution of the reconstruction procedure.) Our surface detection algorithm [12] has been applied to the reconstructed array and it produced an estimated surface of the cast consisting of a connected collection of 12,176 faces of voxels. This collection is displayed in Figure 2 using our variant of Warnock's shading rule. As can be seen, the small square surfaces cause an undesirable visual effect. Figure 3 shows the display shown in Figure 2 after our smoothing procedure has been applied with W = 0.8. (All our displays have been photographed from the screen of a Ramtek display system using a 240 x 240 raster array.)

In spite of the very large number of surface elements, our method required only 1 minute for hidden surface removal and shading, resulting in Figure 2, and a further 8 seconds for smoothing and displaying, resulting in Figure 3. (All timings reported in this note, except where otherwise stated, refer to CDC3500 CPU time and FORTRAN coding.)

We applied the hidden surface removal and shading method of Greenleaf et al [9] to the same object (this method involves sorting the faces according to distance from the observer), and we

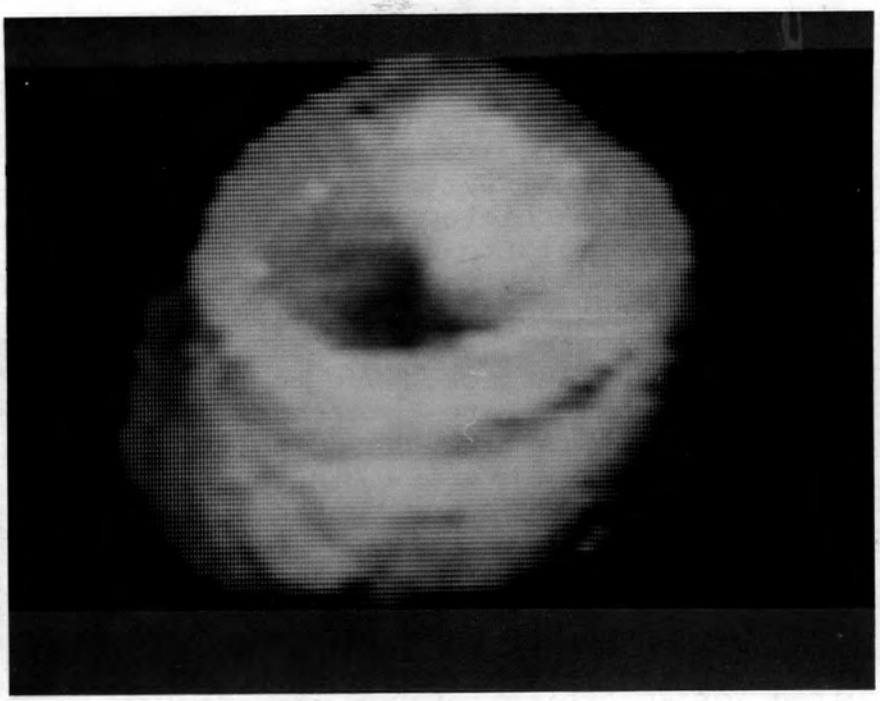


Figure 1

A television image of the plastic cast of a canine left ventricle.

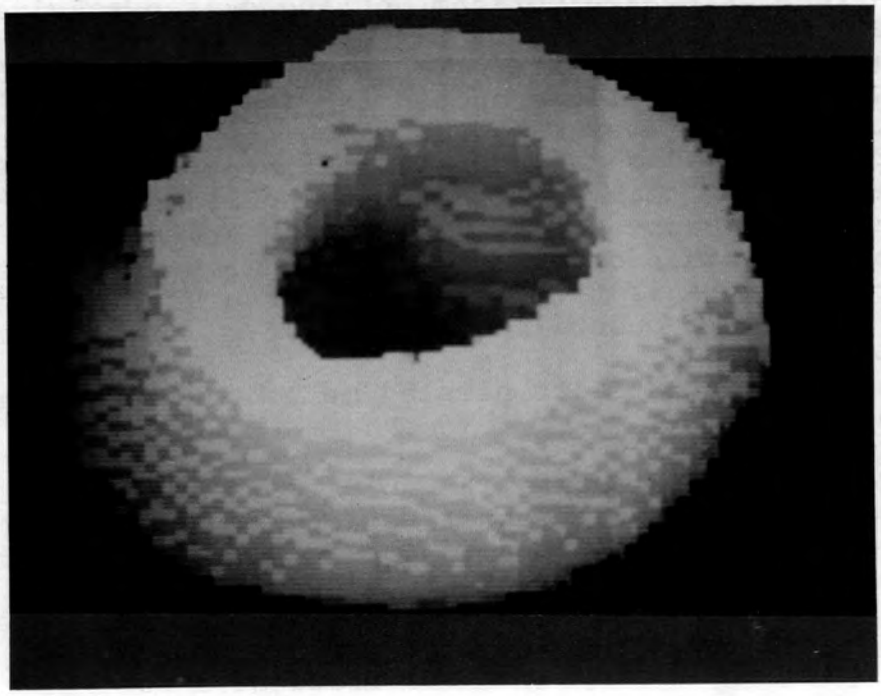


Figure 2

Computer-generated display of the cast which was determined from 28 reconstructed cross-sectional levels of the cast.

found that it needed over 5 minutes.

We also applied the method described by Gouraud [7], which incorporates smoothing into the hidden surface removal and shading process. This approach modifies the computation of the shading on each surface so that continuity exists across surface boundaries. This continuity can be achieved by assigning as normal at a vertex the average of the normals to each surface associated with this particular vertex. Each surface has a different shading for each of its vertices and the shading at any particular point inside the surface has to be computed by linear interpolation of the shading at the vertices of the surface. Due to the large number of the surfaces in the objects we are interested in (typically over 10,000), this approach is time consuming. In our implementation, Gouraud's method required over 16 minutes. Also, the results are far from acceptable, because the surfaces are either parallel or perpendicular to each other. Figure 4 shows a display produced by Gouraud's method. While it is the case that changes to Gouraud's method which produce better quality images have been reported on (e.g., [14]), these changes further increase the already considerable computer time requirement. Note, in particular, that the time required for smoothing in such methods increases with the complexity of the object to be displayed (e.g., normals have to be averaged at each of the vertices of faces), while the time required for our simple low-pass filter depends only on the size of the raster display matrix.

Our timing for Gouraud's algorithm is consistent with what has been previously reported in the literature. Watkins [16], on whose hidden surface removal algorithm Gouraud's is based, reported that it took 2.5 minutes to produce and display on a PDP-10 Figure 35 of his paper, which consisted of only 345 blocks, six polygons per block. Note also the times quoted in Figure 4 of Gouraud [7]; over 1.5 minutes for surfaces based on a few polygons. More recently, Blinn and Newell [1] reported that images of objects made up from 26 bicubic patches took about 25 minutes to display on a PDP-11/45. It does therefore appear that our methods are at least an order of magnitude faster than other more general methods described in the literature would be if they were applied to surfaces made up from numerous faces of a cuberille.

We now demonstrate our display method on two actual reconstructed organs.

In Figure 5, we display the surface of an isolated canine heart whose 64 x 64 x 30 reconstruction has been produced from 50 x-ray projections. The detected surface consists of 13,650 elements and the figure required under 1.5 minutes for hidden surface removal, shading, smoothing and displaying.

In Figure 6, we display the surface of a part of the left lung of a dog. Thirty-five x-ray projections of an intact dead dog have been taken and the thorax and contents of the dog were reconstructed as a 64 x 64 x 64 array. The input to the surface detection algorithm was a 28 x 64 x 64 subarray containing most but not the whole of the left lung. The surface detection algorithm [12] was applied to detect the surface of the lung in the thorax (11,500 surface elements) and the result is displayed in Figure 6. The imprints in the lung of the heart and of the major airway about it are clearly visible. This figure required 1 minute for hidden surface removal, shading, smoothing and displaying.

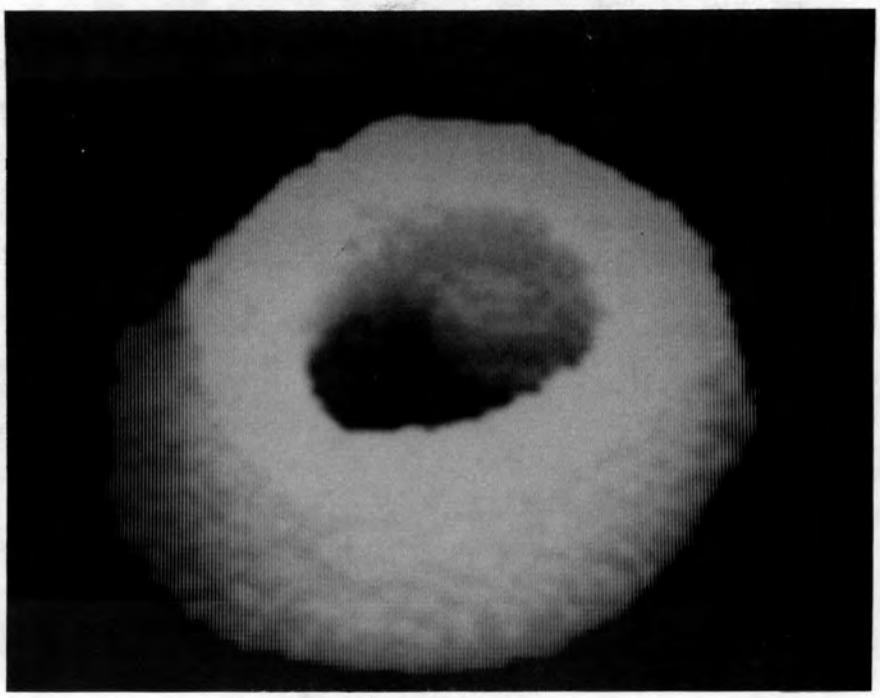


Figure 3

Computer-generated display of the cast smoothed by the proposed low-pass filter procedure.

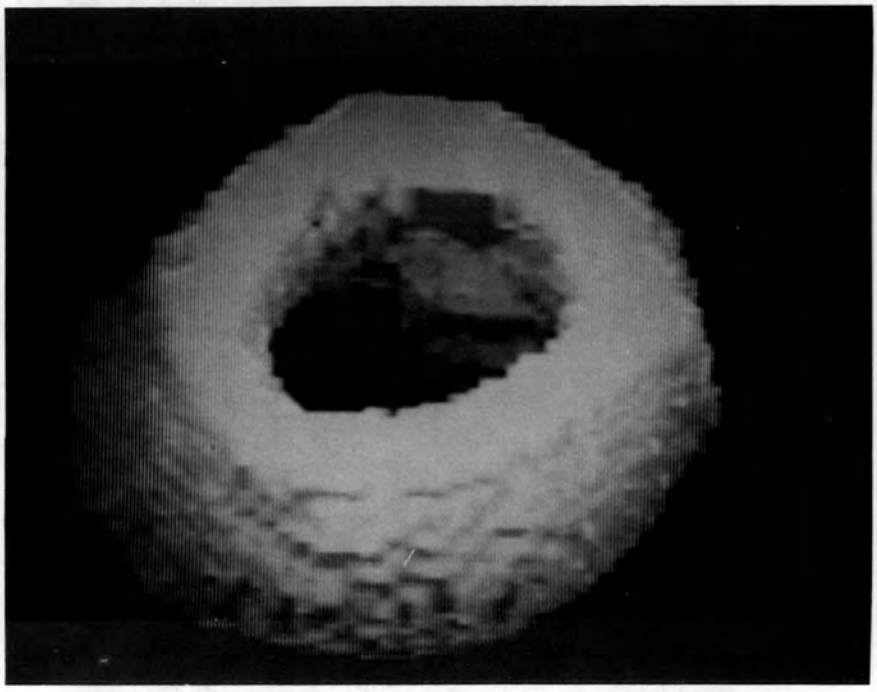


Figure 4

Computer-generated display of the cast by the use of Gouraud's shading procedure.

Figure 5

Computer-generated display of a canine heart which was determined from 30 reconstructed cross-sectional levels of the heart.

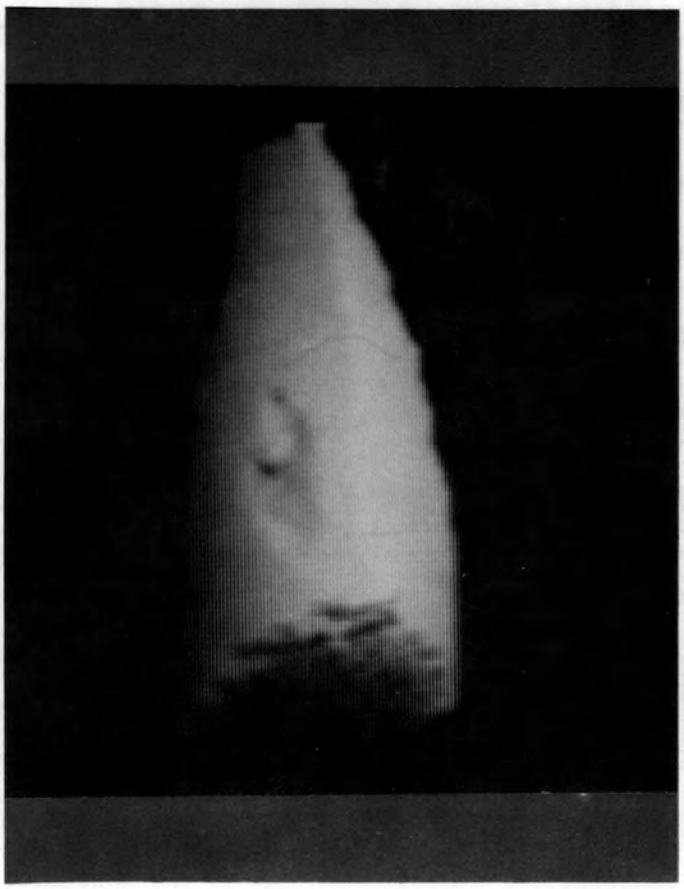


Figure 6

Computer-generated display of a part of a canine left lung which was determined from 64 reconstructed cross-sectional levels of the intact thorax of a dead dog.

DISCUSSION

There are many different reasons for displaying three-dimensional surfaces on a cathode ray tube. The relative importance of the quality of the display and the cost of producing the display depend on the purpose for which the display is being created.

For example, in designing a vase using computer graphics, the quality is of extreme importance, since esthetic appreciation of what the real product would look like can only come from a very realistic display. The cost of display is of lesser importance, since under any circumstances it is likely to be only small part of the total manufacturing cost.

On the other hand, when images are to be displayed for the purpose of medical diagnosis, it is the usefulness of the image and not its realism which is important. For example, ultrasound images often look very different from the organs they image, but nevertheless contain essential information for making a correct diagnosis. In a typical application for our purposes, we desire to provide the surgeon with the gross structure of a deformed heart to avoid any potentially disastrous surprises during an operation, but we need not give him an artistic rendering of his patient's innards. On the other hand, reduction in cost is essential: computer time for the display will have to be borne by the patient.

Another difference between the two situations described above is the importance of quantitative aspects. The shape and even size of a vase can and usually will be changed for the sake of better appearance. Correct reproduction of shape and size of organs may be of great significance for diagnostic purposes.

The methods that have been proposed in the last section are not intended to compete with the techniques that have been developed in recent years for the purpose of realistic surface modelling by computer [1,2,4,14,15]. For example, they are not suitable for showing specular reflections on the surface. On the other hand, they are extremely fast and quantitatively accurate.

A point of general interest here is the speed of the display in spite of the large number of faces involved. It is commonly believed that increasing the number of polygons making up the surface would increase the time for hidden surface removal (see, e.g., [14]). Our work has demonstrated that, provided the larger number of polygons are simpler in shape and degrees of freedom of orientation, reduction in computer time is possible. In spite of the restrictions on shape and orientation, due to the smaller size of the polygons, the quantitative accuracy of the surface is likely to be better than what can be achieved with a significantly smaller number of more general surface elements. Another advantage of using small surface elements is that the images show no observable Mach band effects, especially after smoothing.

Some comments are in order about our smoothing method. It is an example of a low-pass filter. Such filters are commonly used in picture processing for the elimination of high frequency components (such as sharp edges and isolated dots). Nevertheless, they have been either totally ignored in the literature on the display of three-dimensional surfaces, or have been discarded without much ado. ("A brief experiment with this scheme proved unencouraging." says Crow [4] about one such filter.)

We believe that for certain types of surfaces (including the ones we are interested in), smoothing by a low-pass filter, such as the one we have suggested in the last section, is a very appropriate way of improving image quality. First of all, such smoothing is extremely inexpensive. In particular, the cost depends only on the size of the raster display matrix, and not on the complexity of the object to be displayed. (Smoothing for each of the 240 x 240 displays in Figures 3, 5, and 6 took 8 seconds of CDC 3500 CPU time when programmed in FORTRAN.)

Secondly, smoothing this way gets rid of many of the artifacts that one is worried about in three-dimensional display of surfaces, especially if the surface elements are likely to occupy areas on the display screen which are about the same size as the smoothing matrix (3 x 3 in the case described in the last section). For example, our method smooths the silhouette of the object automatically. This is especially appropriate when organs are displayed against a dark background, since the surface is likely to be curving away from the light source near the edge. (Compare Figure 2 with its smoothed version in Figure 3.) More generally, even though our method does not make use of partial areas (in the presmoothed picture each raster point is determined by one and only one surface), the contribution of neighboring raster points to a given raster point has an effect usually similar to what would be achieved by explicitly incorporating partial areas into our calculation.

There is one undesirable consequence of the way we produce our unsmoothed shaded picture which cannot be compensated for by any amount of smoothing. It is possible that a "visible" surface element is positioned so that it is missed by all dots on the display screen. (This results in "aliasing", see e.g., [1,4].) We can reduce the frequency of occurrence of this problem, and, at the same time, improve the general appearance of our displays, by the following procedure, which corresponds to a low-pass filtering of the object to be displayed rather than just the display itself (cf., [4]).

For the purpose of hidden surface removal and shading we may assume an array of display dots which is finer than the actual raster array. Typically, we may use four or even nine dots for each display point. Having done the hidden surface removal and shading, we can apply the smoothing algorithm to calculate the smoothed value at the actual display points based on the larger array of unsmoothed values. (The smoothing matrix will usually require a larger number of elements than what we have used in the last section. In general, the number of points in the smoothing matrix should be approximately the same as the number of display dots needed for a typical surface element.)

This approach will capture faces missed by our earlier approach, and it will generally produce a better quality picture, because of its more correct treatment of partial areas. However, it slows down the process (proportionately to the number of dots we use for each raster point) and it may require excessive storage. For example, if for a 240 x 240 raster display we wish to use four dots for each roster point and 256 grey levels, and if we believe that 256 levels in the Z-buffer are enough to make correct decisions in the hidden surface algorithm (as is likely to be the case when displaying an organ using faces in a cuberille), then the

number of 16 bit words needed for the hidden surface removal and shading is 230,400. Since the brightness and distance from observer for nearby display dots is likely to be very similar, simple offset coding can reduce the storage requirement considerably, but only at the expense of time that has to be spent on decoding during the hidden surface removal and shading process.

From the above discussion it appears that considerable improvement to the quality (from the point of view of realistic appearance) of the pictures produced by our method can be achieved by more expensive methods based on the same principles.

CONCLUSIONS

We have proposed methods for hidden surface removal, shading and smoothing which are very efficient if the surface to be displayed consists of faces of a cuberille. Such surfaces are appropriate as approximations to organ surfaces, especially in the context of computerized tomography. Applied to such surfaces, our algorithms produce displays which are as good or better than displays produced by previously proposed more general methods, but in a small fraction of the time required by those methods.

ACKNOWLEDGEMENTS

The authors wish to thank members of the Biodynamical Research Unit at the Mayo Clinic, especially Dr. James F. Greenleaf for valuable discussions and Drs. Lowell D. Harris, Erik L. Ritman and Richard A. Robb for providing test data. We are also grateful to members of the Medical Image Processing Group at the State University of New York at Buffalo, especially to Dr. Arnold Lent for valuable discussions and Ms. Mann-May Yau for pointing out possible improvements to an earlier version.

REFERENCES

1. J. F. Blinn and M. E. Newell, "Texture and reflection in computer generated images," Comm. ACM 19, 10, 542-547 (Oct. 1976).

2. E. Catmull, "Computer display of curved surfaces", Proc. Conf. on Comptr. Graphics, Pattern Recognition, and Data Structure, (published by IEEE), May 1975, pp. 11-17.

3. J. H. Clark, "Hierarchical geometrical models for visible surface algorithms," Comm. ACM 19, 10, 547-554 (Oct. 1976).

4. F. C. Crow, "The aliasing problem in computer-synthesized shaded images," Tech. Rep. UTEC-CSC-76-015, Dep. Comptr. Sci., U. of Utah, Salt Lake City, Utah, March 1976.

5. R. Gordon and G. T. Herman, "Three-dimensional reconstruction from projections: a review of algorithms," International Review of Cytology 38 (eds.: G. H. Bourne, J. F. Danielli) Academic Press, New York, NY, pp. 111-151, 1974.

6. R. Gordon, G. T. Herman, and S. A. Johnson, "Image reconstruction from projections," Scientific American 233, 4, 56-68, (1975).

7. H. Gouraud, "Continuous shading of curved surfaces," IEEE Trans. Computers C-20, 6, 623-629 (June 1971).

8. J. F. Greenleaf, E. L. Ritman, C. M. Coulam, R. E. Sturm, and E. H. Wood, "Computer graphic techniques for study of temporal and spatial relationships of multidimensional data derived from biplane Roentgen videograms with particular reference to cardioangiography," Comptrs. Biomedical Research 5, 368-387, 1972.

9. J. F. Greenleaf, J. S. Tu, and E. H. Wood, "Computer generated three-dimensional oscilloscopic images and associated techniques for display and study of the spatial distribution of pulmonary blood-flow," IEEE Trans. Nuclear Sci. NS-17, 3, 353-359 (June 1970).

10. G. T. Herman, "A relaxation method for reconstructing objects from noisy x-rays," Math. Programming 8, 1, 1-19 (Jan. 1975).

11. G. T. Herman, and H. K. Liu, "Dynamic boundary surface detection", Comptr. Graphics Image Processing (in press).

12. H. K. Liu, "Two-and three-dimensional boundary detection," Comptr. Graphics Image Processing 6, 2, 123-134 **(April 1977)**.

13. J. C. Mazziotta, and H. K. Huang, "THREAD (Three-dimensional Reconstruction and Display) with biomedical applications in neuron ultrastructure and computerized tomography," Proc. Nat. Comptr. Conf., (published by AFIPS), pp. 241-150 **(June 1976)**.

14. B. T. Phong, "Illumination for computer generated pictures," Comm. ACM 18, 6, 311-317 (June 1975).

15. B. T. Phong, and F. C. Crow, "Improved rendition of polygonal models of curved surfaces," Proc. 2nd USA-Japan Comptr. Conf. (published by AFIPS) 475-480 (August 1975).

16. G. W. Watkins, "A real-time visible surface algorithm," Tech. Rep. UTEC-CSC-70-101, Dep. Comptr. Sci., U. of Utah, Salt Lake City, Utah (June 1970).

APPENDIX

The ordering of square surface elements.

In this appendix we prove the theorem in the section entitled "Methods".

Much of what is done in the Appendix is applicable to a generalization of the concept of a cuberille to arbitrary polyhedrons in place of the equal sized cubes. We call this generalized concept a "dissection" of the three-dimensional space.

Definition. A dissection (of three-dimensional space) S is a pair $\{\Phi, C\}$, where

(i) Φ is an infinite collection of simple closed convex polyhedrons (called the components of Φ), which cover the whole three-

dimensional space and which are such that if two elements of Φ intersect then the intersection is a face for both components, and

(ii) C is a mapping which associates with every face a point in the interior of the face.

Definition. We say that dissection $S = \{\Phi, C\}$ is locally ordered if the following condition is satisfied.

Whenever,

(i) O is a point,

(ii) ϕ is a component of Φ

(iii) α and β are faces of ϕ,

(iv) A is a point on α but not on β,

(v) B is a point on β but not on α,

(vi) O, A and B are collinear, with A and B on the same side of O.

(vii) $0 < |OA| < |OB|$,

then

(viii) $|OC(\alpha)| < |OC(\beta)|$.

Lemma 1. A cuberille is locally ordered.

Proof. Let S be a cuberille, O be an arbitrary point, ϕ be a component of Φ, α and β be faces of ϕ, A be a point on α but not on β and B be a point on β but not on α, such that O, A and B are collinear and $0 < |OA| < |OB|$. We distinguish between two cases.

Case (i) α and β are parallel.

Let ℓ be the line through $C(\alpha)$ and $C(\beta)$ and let O' be the unique point on ℓ such that the line segment OO' is perpendicular to ℓ. Clearly, OO' is parallel to α and β, and so $0 < |O'C(\alpha)| < |O'C(\beta)|$. But

$$|OC(\alpha)|^2 = |OO'|^2 + |O'C(\alpha)|^2,$$

$$|OC(\beta)|^2 = |OO'|^2 + |O'C(\beta)|^2,$$

and so

$$|OC(\alpha)| < |OC(\beta)|.$$

Case (ii) α and β are perpendicular with an edge in common.

First note that from the fact that O, A and B are collinear and $0 < |OA| < |OB|$, it follows that O is on the opposite side of α than $C(\beta)$ is and is on the same side of β as $C(\alpha)$ is.

Let ℓ be the common edge of α and β and let P be the unique plane perpendicular to ℓ going through the points $C(\alpha)$ and $C(\beta)$. Let O' be the unique point on P such that the line segment OO' is perpendicular to P. Since OO' is parallel to both α and β, O' must be on the opposite side of α than $C(\beta)$ is and on the same side of β as $C(\alpha)$ is. Observing Figure 7 we see that

$$|O'C(\alpha)|^2 = a^2 + r^2 - 2ar\cos(\theta - \pi/2),$$

$$|O'C(\beta)|^2 = a^2 + r^2 - 2ar\cos\theta.$$

where $\frac{\pi}{2} < \theta \leq \pi$. Hence it follows that $|O'C(\alpha)| < |O'C(\beta)|$. But

$$|OC(\alpha)|^2 = |OO'|^2 + |O'C(\alpha)|^2,$$

$$|OC(\beta)|^2 = |OO'|^2 + |O'C(\beta)|^2,$$

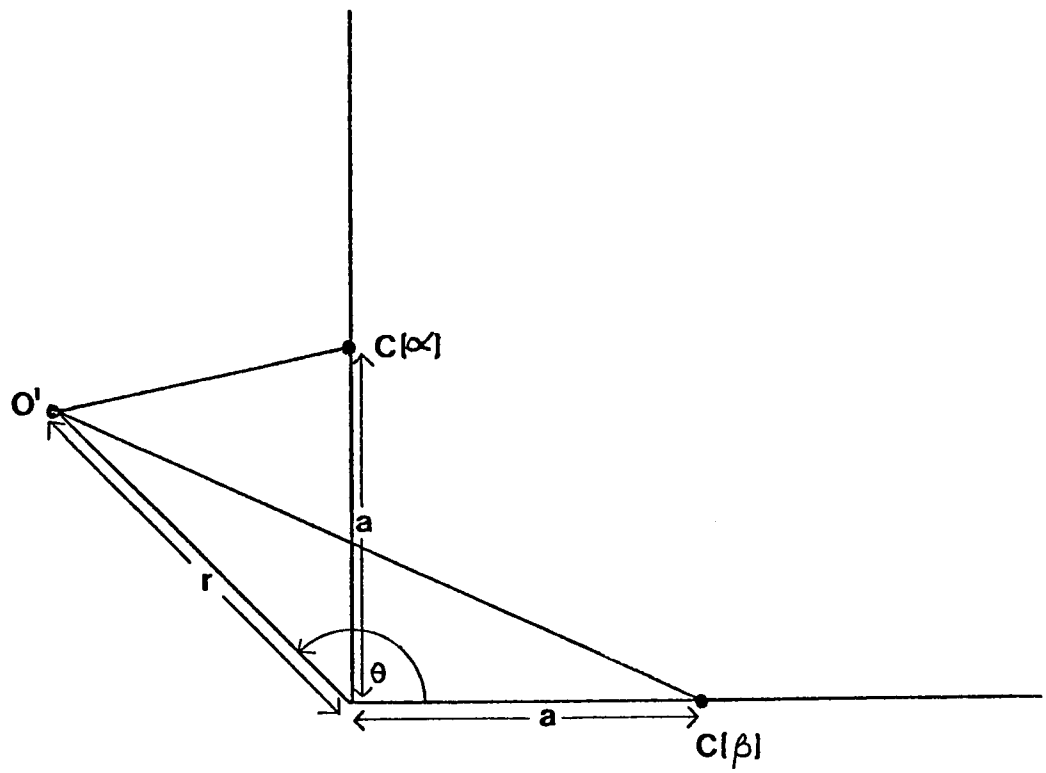


Figure 7

Perpendicular faces of a cuberille.

and so

$$|OC(\alpha)| < |OC(\beta)|.$$

<u>Definition.</u> We say that a dissection $S = \{\Phi, C\}$ is <u>globally ordered</u>, if the following condition is satisfied.

Whenever

(i) O is a point,
(ii) α and β are two different faces of S, such that O does not lie in the plane of either α or β,
(iii) A is a point in the interior of α,
(iv) B is a point in the interior of β,
(v) O, A and B are collinear, with A and B on the same side of O,
(vi) $0 < |OA| < |OB|$,

then

(vii) $|OC(\alpha)| < |OC(\beta)|$.

<u>Lemma 2</u>. Every locally ordered dissection is globally ordered.

<u>Proof</u>. Let S be a locally ordered dissection, O be a point, α and β be two different faces of S such that O does not lie in the plane of either α or β, A be a point in the interior of α and B be a point in the interior β such that O, A and B are collinear, with A and B on the same side of O and $0 < |OA| < |OB|$.

We are going to define sequences $A_0, A_1, \ldots, A_n$ of points, $\alpha_0, \alpha_1, \ldots, \alpha_n$ of faces and $\phi_1, \ldots, \phi_n$ of components which have the following properties.

(i) $\alpha_0 = \alpha$.
(ii) $\alpha_n = \beta$.
(iii) For $1 \leq i \leq n$, α_{i-1} and α_i are faces of ϕ_i.
(iv) For $1 \leq i \leq n$, A_{i-1} is a point on α_{i-1}, but not on α_i.
(v) For $1 \leq i \leq n$, A_i is a point on α_i, but not on α_{i-1}.
(vi) For $1 \leq i \leq n$, O, A_{i-1} and A_i are collinear, with A_{i-1} and A_i on the same side of O,
(vii) For $1 \leq i \leq n$, $0 < |OA_{i-1}| < |OA_i|$.

Let us assume for now that we have already succeeded in defining such sequences. Then, for $1 \leq i \leq n$, the fact that S is locally ordered implies that

$$|OC(\alpha_{i-1})| < |OC(\alpha_i)|.$$

This, combined with the fact that $\alpha_0 = \alpha$ and $\alpha_n = \beta$, implies that

$$|OC(\alpha)| < |OC(\beta)| ,$$

which is what we need to prove.

We now give the construction of the sequences $A_0, A_1, \ldots, A_n$ $\alpha_0, \alpha_1, \ldots, \alpha_n, \phi_1, \ldots, \phi_n$, which satisfy the properties (i) - (vii) described above.

First of all, it is easy to see that there exist points A' on α and B' on β such that O, A' and B' are collinear and the line segment A'B' does not cross an edge of any face in the cuberille.

We define $A_0, A_1, \ldots, A_n$ to be the set of all points on the

line segment A'B' which lie on a face of S, arranged in an order so that, for $1 \leq i \leq n$, $|OA_{i-1}| < |OA_i|$. Since O, $A_0 = A'$ and B' are collinear and $0 < |OA'|$, (vi) and (vii) are satisfied.

Since $A_0, A_1, \ldots, A_n$ lie on the line segment A'B' and also on faces of S, it follows by the choice of A' and B' that none of the A_i's lie on an edge of a face of S. Hence, for $0 \leq i \leq n$, there is a unique face α_i such that A_i lies on α_i. Also for $1 \leq i \leq n$, there is a unique component ϕ_i such that α_{i-1} and α_i are faces of ϕ_i. (This is the component which contains the segment $A_{i-1}A_i$.) This completes our definition of the α_i's and the ϕ_i's. Properties (i) to (v) are clearly satisfied, completing our proof of Lemma 2.

Lemma 3. If a dissection $S = \{\Phi, C\}$ is globally ordered then it satisfies the following property. For any point O, for any different faces α and β of S, and for any points A and B such that A is in the interior of α, B is in the interior of β and O, A and B are collinear with A and B on the same side of O (not coinciding with O),

$$|OA| < |OB|$$

if and only if

$$|OC(\alpha)| < |OC(\beta)|.$$

Proof. If $|OA| < |OB|$, then $|OC(\alpha)| < |OC(\beta)|$ by the Lemma 2. If $|OC(\alpha)| < |OC(\beta)|$, $|OA| \leq |OB|$ (by Lemma 2), but $|OA| \neq |OB|$ since then we would have A = B, contradicting the condition that A and B lie in the interior of two different faces.

The theorem in the main body of the paper is now an immediate consequence of Lemmas 1, 2 and 3. The derivation in this Appendix was given in such a way that generalization to all locally ordered dissections is immediate.

DISCUSSION:

Pizer asked about the time for the program to display the picture after the surfaces had already been identified, what the resolution of the display was and the number of gray scales used. Liu answered that it took about one minute on the CDC 8500 after surface identification and removal of hidden surfaces, most of which was disk I/O time, to display the picture in 256 gray levels with a display resolution of 240x240.

Porath wondered if the radiation dose had been estimated for a 4-dimensional reconstruction of the heart and Herman replied that he did not know the exact figures but referred to an article to be published in the Mayo Clinic Proceedings. Pullan stated that the dosage might be as high as 50 rads for an investigation like this and that in real clinical scans there is a very small difference in attenuation coefficients between blood in the heart and the heart muscle and that you must use even higher radiation dose to visualize that difference. Herman replied that this was not a standard CT-machine. The machine which is to be built at the Mayo Clinic will have 28 X-ray sources and opposed image intensifiers in a semicircle, electronically switched, each being on for 1/100 th second. Furthermore, contrast injection is going to be used.

Pullan then asked if they were capable of collecting all the data while the bolus was going through the heart or if they were working at an equilibrium. This question was answered by Coulam, who said that theoretically one can collect all the data as the bolus passes the heart, but they have been working with brominated fluorocarbides as a blood analog. This was preferable due to its very slow removal from the circulation (1/2-1 hour). He also made the comment that they gather the data during roughly 2 complete heart beats and that they delivered 2 rads/second to the animal. He also mentioned that conventional processing could take up to two weeks to get the different cross-sections and that they have to go to hard-wired reconstructors, etc. Herman added that the temporal resolution was 1/60 th second and that the data was filtered over time as well as over space and that the detection of moving surface is therefore fairly accurate.

Rogers asked what surface detection algorithm had been used, and Herman described it as a gradient search procedure which builds up a surface from an interactively chosen point near the surface. The algorithm has also a capability of back-tracking if it is lost.

Rogers wondered also if the operator handled problems of discontinuities such as at the edge of cut and Herman replied that the operator specifies the plane of cut and the surface then follows that plane.

Gruber asked if the surface was constructed from edges of different planes, and Herman explained that this was not the case; instead, they constructed a continuous surface from the square surfaces of the "voxels."

Todd-Pokropek stated that he was surprised that they did not use the technique of patches since he found the cubic structure of the picture disturbing. Herman replied that it was more natural, in CT-scans, to display the voxels as they are and also that the hidden surface removal and shading is faster than for more general surface. Liu emphasized that it would be slower to work with patches in this application.

Metz made the comment that his group obtained the illusion of depth in the display by modulating the brightness by the cosine of the angle and by the distance to the observer and asked if they also used some shadowing since the inside of the heart was dark. Coulam explained that this probably was caused by the fact that when planes approach a parallel orientation to the light source, they get darker and darker.

INTERACTIVE BIOMEDICAL IMAGE PROCESSING*

E. L. Hall and R. C. Gonzalez
Departments of Electrical Engineering
and Computer Science
The University of Tennessee
Knoxville, Tennessee 37916

ABSTRACT

Interactive image processing techniques are useful in medical applications in which the solution depends upon the subject or organ examined and in cases in which total automation is too costly or difficult to implement. A survey of interactive image processing techniques is presented in this paper. Each of the techniques presented has been or could easily be implemented in hardware using presently available microcomputers and operating at real time TV rates. The selection of interactive capabilities is briefly considered. Interactive image enhancement using histogram specification, spatial filtering, and pseudocolor mapping are presented. Examples from radiography, mammography, and computer tomography are given. This comparative study illustrates the importance and availability of interactive devices for improving patient care through image processing.

INTRODUCTION

The advantage of interactive processing is that the scientist has direct control and knowledge not only of the input data but also of intermediate calculations and results. The main drawback is that operating a system based on this approach requires an expert who can recognize a significant result.

Medical imaging techniques are in routine use in radiology, for example, scintigraphy, radiography, computer tomography, and ultrasonography. A recent survey of display systems [1] indicates a number of limitations of current displays. The purpose of this paper is to examine several interactive features which could be added to a medical display.

A selected set of implementable, interactive image enhancement techniques is described below. Since the interactive requirements vary with the examination and subject, the "menu design" approach is first considered. The histogram of an image is an easily computed statistic which provides a large amount of information about an examination. Histograms of chest radiography are presented which demonstrate the relationship between a properly exposed radiograph and its histogram. It is hypothesized that for each examination and major subject

*Work partially supported by NASA Contract NAS8-29271.

category, an ideal underlying histogram exists. Histogram specification is then presented which shows that, given the underlying distribution or parameters which specify it, it is possible to transform a given image into one with the specified histogram. The application of this technique to image enhancement follows directly. The interactive feature of the implementation permits a single machine to be used for a variety of examinations and subject ranges by the selection of the appropriate distribution.

Spatial frequency techniques are widely used for image enhancement; however, the shape of the spatial filter varies considerably depending upon the examination and subject. Edge enhancement or spatial highpass filtering can be implemented using a local unsharp masking operation. This technique is demonstrated for improving the visibility of fine detail in a radiograph. However, edge enhancement is not desirable in all examinations and often exactly the opposite technique, lowpass spatial filtering, is required. This effect is often desirable in nuclear medicine imaging for noise smoothing and in radiography for demonstrating mass lesions. An example, xeroradiogram which has inherent edge sharpening is presented in which the visibility of a mass lesion can be improved by lowpass spatial filtering.

Since the human visual system has highest acuity in the fovea, which contains color receptors, the use of pseudocolor mapping offers a potential improvement for image enhancement applications. However, the mapping from a black-white, one-dimensional space, to a chromatic, three-dimensional space, is difficult to specify. One set of solutions is to specify a curve path such as the spectral colors or the chromaticity curve of a blackbody radiator. The interactive approach would indicate the use of a set of curve paths which could be selected easily from the display. An example of pseudocolor mapping for a computer tomography brain image is presented to demonstrate the enhanced visibility of a lesion by means of a pseudocolor presentation.

SELECTING INTERACTIVE OPTIONS

An important consideration in any interactive application is the selection of a set of interactive options. This step is often called "menu design." Although only the options which are placed on the menu can be selected, it often appears that the difficulty of implementating the options has resulted in sparce selections. Due to the infinite variety of possible options, attention will be focused on image enhancement techniques which perform a necessary function and for which an interactive solution is required because of the dependency of the solution on the examination and subject.

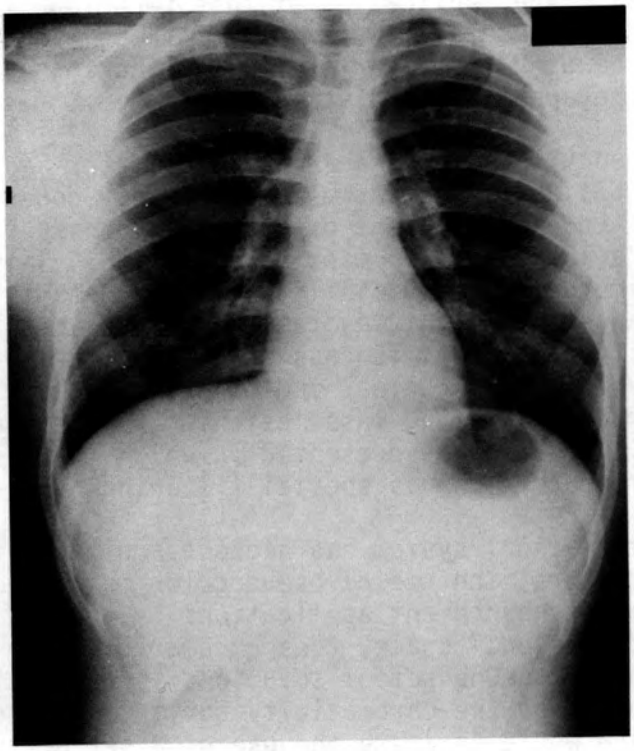
Among the many techniques which fall into the category the following are representative:

(1) Histogram specification,
(2) Spatial frequency filtering,
(3) Pseudocolor mapping.

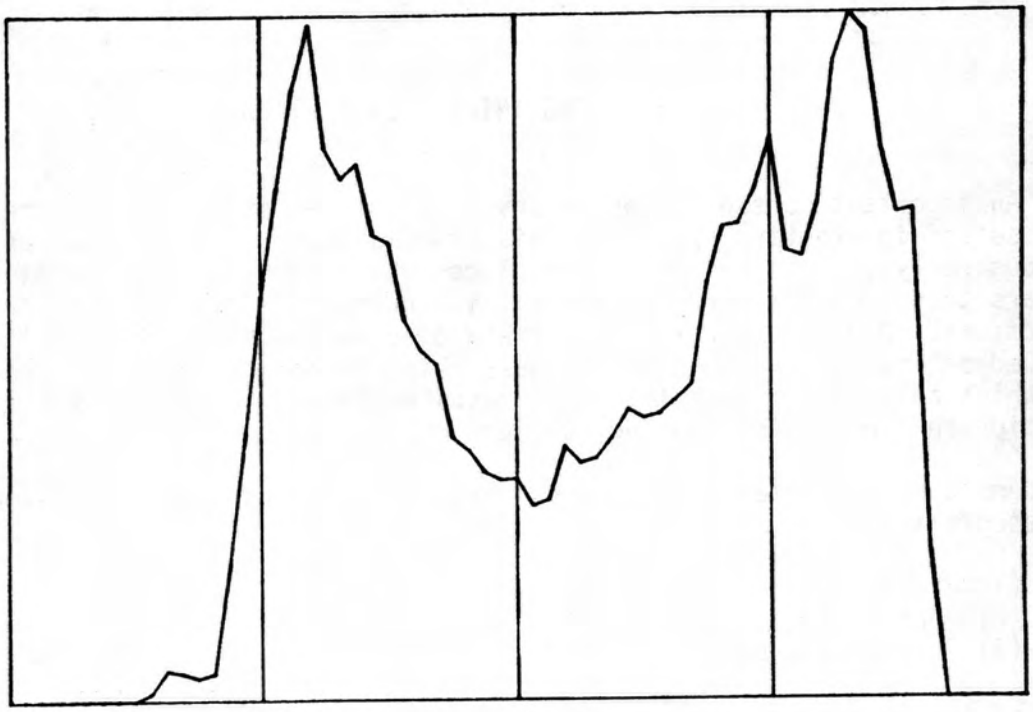
Each of these techniques is described in the following sections.

HISTOGRAM SPECIFICATION

For any radiographic image, $f(x, y)$, a histogram is easily computed by simply tabulating the number of occurrences of each gray level. The histogram

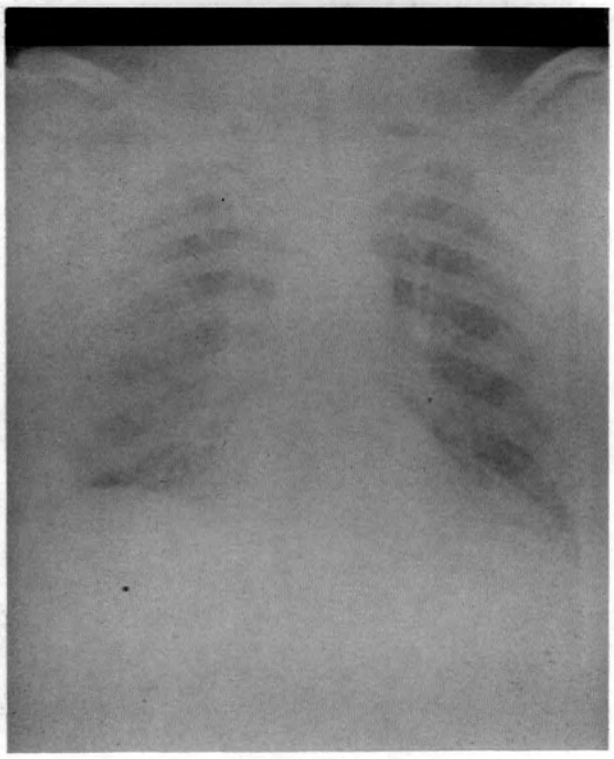


(a)

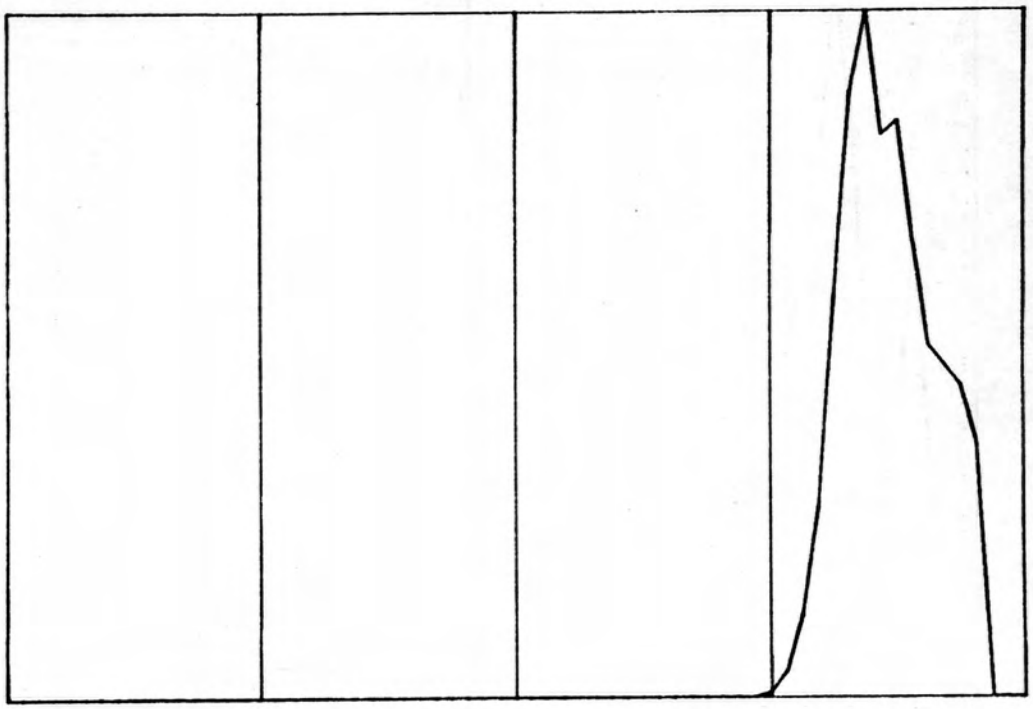


(b)

Figure 1. (a) Properly exposed chest radiograph of adult. (b) Corresponding histogram.

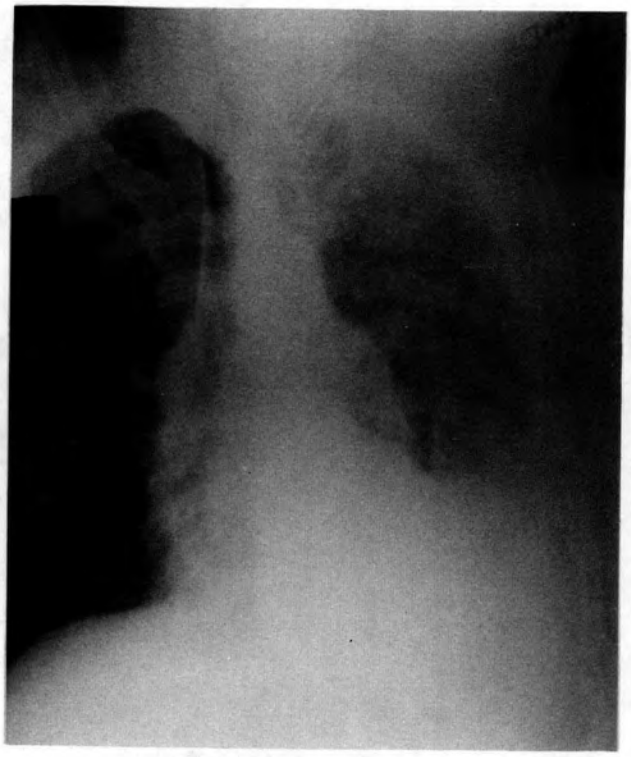


(a)

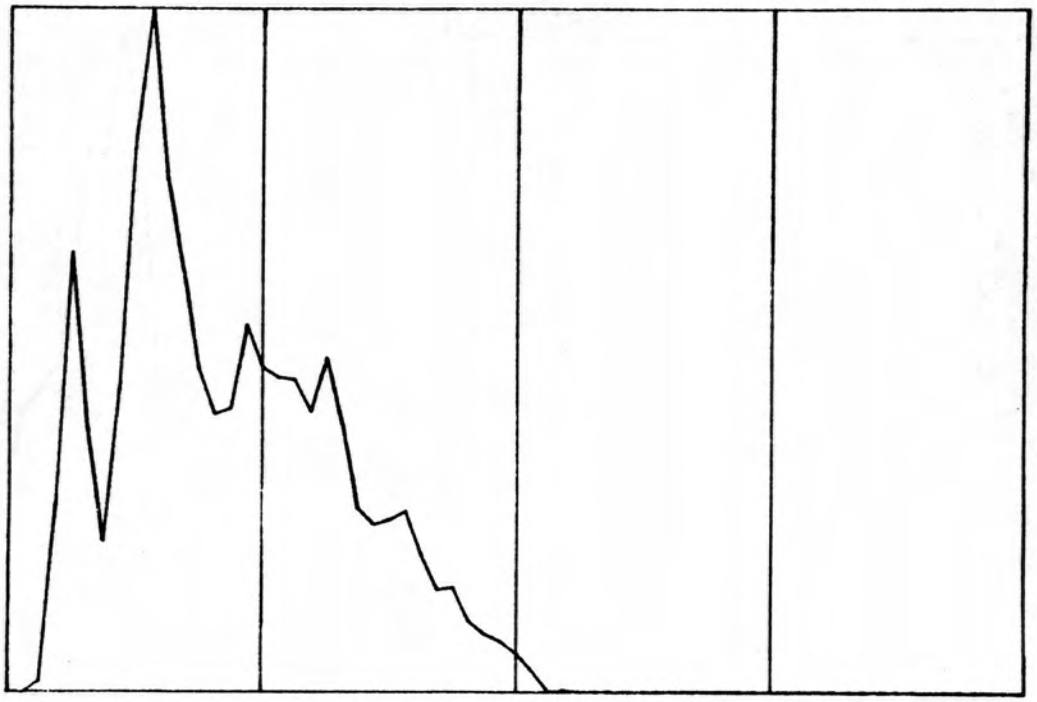


(b)

Figure 2. (a) Underexposed chest radiograph of adult. (b) Corresponding histogram biased to white gray values.



(a)



(b)

Figure 3. (a) Overexposed and fogged radiograph of adult. (b) Corresponding histogram biased to dark gray values.

provides a large amount of information about film quality and often the pathology shown by the examination. To provide maximum benefit, the histogram must be made specific to each type of examination, and ranges of anatomical variation must also be imposed. For example, a properly exposed chest radiograph of an adult and the corresponding histogram are shown in Figure 1. Note the wide contrast range which is characteristic of the wide latitude of a proper exposure. The two major peaks are characteristic to the anatomy, with the dark peak corresponding to the lungs and air space and the white peak corresponding to the dense bone structure. These results may be contrasted with the underexposed radiograph and its histogram shown in Figure 2. Note that the washed out appearance of the radiograph is clearly reflected in the single peak and narrow range of the histogram. As another example, consider the overexposed and fogged radiograph shown in Figure 3. The corresponding histogram reflects the unusual dark values in the film by showing a narrow band of values biased to the lower gray values. These examples illustrate that for a given examination and subject range there may be an underlying histogram which could be expected from a properly administered exam. This hypothesis may be considered equivalent to assuming an underlying probability distribution for the gray levels with the parameters of the distribution dependent on the types of examinations and subjects.

We now describe a method [2] for interactively specifying the parameters of the underlying distribution and develop a transformation which converts an image into one which has the specified distribution. The method could be directly applied to the enhancement of improperly exposed films and thus perhaps avoid the necessity of repeated patient exposures.

Suppose that an image to be processed is digitized to form an N X M array of elements (pixels), and let x be a variable which represents the gray level of each pixel. It is assumed for simplicity that x has been normalized to the interval $0 \leq x \leq 1$, where 0 denotes black and 1 denotes white in the gray scale. Attention will be focused in the following sections on transformations of the form

$$y = T(x), \; 0 \leq x \leq 1 \tag{1}$$

which map a gray level x into a level y. The function T(x) is assumed to satisfy the conditions

(a) T(x) is single-valued and strictly monotonic in the interval $0 \leq x \leq 1$,

and

(b) $0 \leq T(x) \leq 1$ for $0 \leq x \leq 1$. (2)

The monotonicity condition preserves the order from black to white in the gray scale of the enhanced image. In addition, condition (b) guarantees a mapping that will be consistent with the allowed range of pixel values.

The gray-level mapping method developed in this section is based on transforming the probability density function of the gray levels in an image to be enhanced. The density function of the levels in the original image will be denoted by $p_x(s)$, while $p_y(y)$ will be used to denote its counterpart in the enhanced image. Although these quantities are in reality discrete functions for a digitized image, the following development will be carried out in continuous form to simplify the explanation. The discrete equivalents are obtained from these results without difficulty.

The functions $p_x(x)$ and $p_y(y)$ are of fundamental importance in describing the visual characteristics of the original and enhanced images. For example, the average brightness of the original image is given by

$$\bar{x} = \int_0^1 x p_x(s) dx \tag{3}$$

while its variance (which is a measure of contrast) is given by

$$\sigma_x^2 = \int_0^1 (x-\bar{x})^2 p_x(x) dx \tag{4}$$

For a given image, we are interested in obtaining an enhanced image with a density function $p_y(y)$ which has been interactively specified by the operator.

The transformation function $y = T(x)$ and its inverse $x = T^{-1}(y)$, are guaranteed to be strictly monotonically increasing in the interval [0, 1] by the conditions set forth in Eq. (2). Under these conditions, $p_y(y)$ can be written in terms of $p_x(x)$ and $T(x)$, as follows [4]:

$$p_y(y) = \left[p_x(x)\frac{dx}{dy}\right]_{x=T^{-1}(y)} \tag{5}$$

The function $p_x(x)$ is obtained from the original image, $p_y(y)$ is specified, and the problem is to determine the transformation function $T(x)$ which will yield the desired $p_y(y)$.

A popular technique, known as histogram equalization [3,4,5], is obtained from Eq. (5) by using the transformation function

$$y = T(x) = \int_0^x p_x(s) ds \qquad 0 \leq x \leq 1 \tag{6}$$

which is the cumulative distribution function of x. The variable "s" in Eq. (6) is a dummy variable of integration. From this equation we have that $dx/dy = 1/p_x(s)$, and Eq. (5) reduces to

$$p_y(y) = 1 \quad 0 \leq y \leq 1 \tag{7}$$

In other words, the use of Eq. (6) yields an image whose gray levels have a uniform density.

Although histogram-equalization can be very useful in some applications, this particular method is not suited for interactive image enhancement since all it can do is produce a density function $p_y(y)$ that is uniform. As will be seen below, however, histogram-equalization can be used as an intermediate step in a transformation which will actually yield a specified $p_y(y)$.

Suppose that the gray levels x of an image are transformed using Eq. (6) to yield a new set of levels z; that is,

$$z = H(x) = \int_0^x p_x(s) ds \quad . \tag{8}$$

From the above discussion, we have that $p_z(z)$ is a uniform density function. If the inverse of Eq. (8), $H^{-1}(z)$, is applied to the z's, we obtain the original image with $p_x(s)$ back. It is noted, however, that if we specify a density function $p_y(y)$ and apply Eq. (6) we would obtain

$$z = G(y) = \int_0^y p_y(s)ds \quad . \tag{9}$$

Although G(y) is in general different from H(x), $p_z(z)$ is the same if either Eq. (8) or Eq. (9) is used. If we now apply the inverse function $G^{-1}(z)$ to the z's, the result would be a set of y levels with the specified density $p_y(y)$. In other words, if the z's have a uniform density, the desired $p_y(y)$ can be obtained by using the inverse mapping $G^{-1}(z)$. Therefore, if the original image is first histogram-equalized and the new (uniform) levels are inverse-mapped using the function $G^{-1}(z)$, the result would be an image whose gray levels have the desired density $p_y(y)$.

The above conclusion can be expressed in terms of a transformation function T(x) from x to y by noting that $y = G^{-1}(z)$ and, since $z = H(x)$,

$$y = T(x) = G^{-1}[H(x)] \quad . \tag{10}$$

The histogram-equalization technique is a special case of Eq. (10) obtained by letting $G^{-1}[H(x)] = H(x)$.

The transformation procedure is summarized in Fig. 4. Figure 4(a) shows the original density function which is calculated from the image to be enhanced. Figure 4(b) shows the transformation function H(x) which maps x into z. Figure 4(c) is the result of this transformation. Figure 4(d) shows the inverse transformation from z to y. The result of this transformation, shown in Fig. 4(e), is the desired probability density function $p_y(y)$. The problem in using Eq. (10) for continuous variables lies in obtaining the inverse function analytically. In the discrete case, this problem is circumvented by the fact that the number of distinct gray levels is usually relatively small, and it becomes feasible to calculate and store a transformation value for each discrete pixel value.

3.1 Interactive Specification of Histograms

Four parameters are used in the histogram specification method described below. The first parameter, denoted by m, specifies a point in the pixel spectrum about which the histogram will be generated. The second parameter, denoted by h, specifies the height of the histogram at the point m. The third and fourth parameters are referred to as the right spread θ_R, and the left spread, θ_L, about m. It is noted that m is related to the mean of the histogram and h to its peakedness. The other two parameters control spread (variance) and symmetry. Although the number of parameters could easily be increased, experiments have indicated that more than four parameters are quite difficult to control interactively, particularly when enhancement speed is an important consideration.

Figure 5(a,b) illustrates the approach for decreasing and increasing the spread about m. Both of these cases involve specifying the point m which can be located anywhere on the interval (0, 1). Then the height is chosen at that point and may assume any value greater than or equal to zero.

For the case of decreasing spread, the parameters θ_R and θ_L specify the angle from vertical, as shown in Fig. 5(a). These parameters can assume values

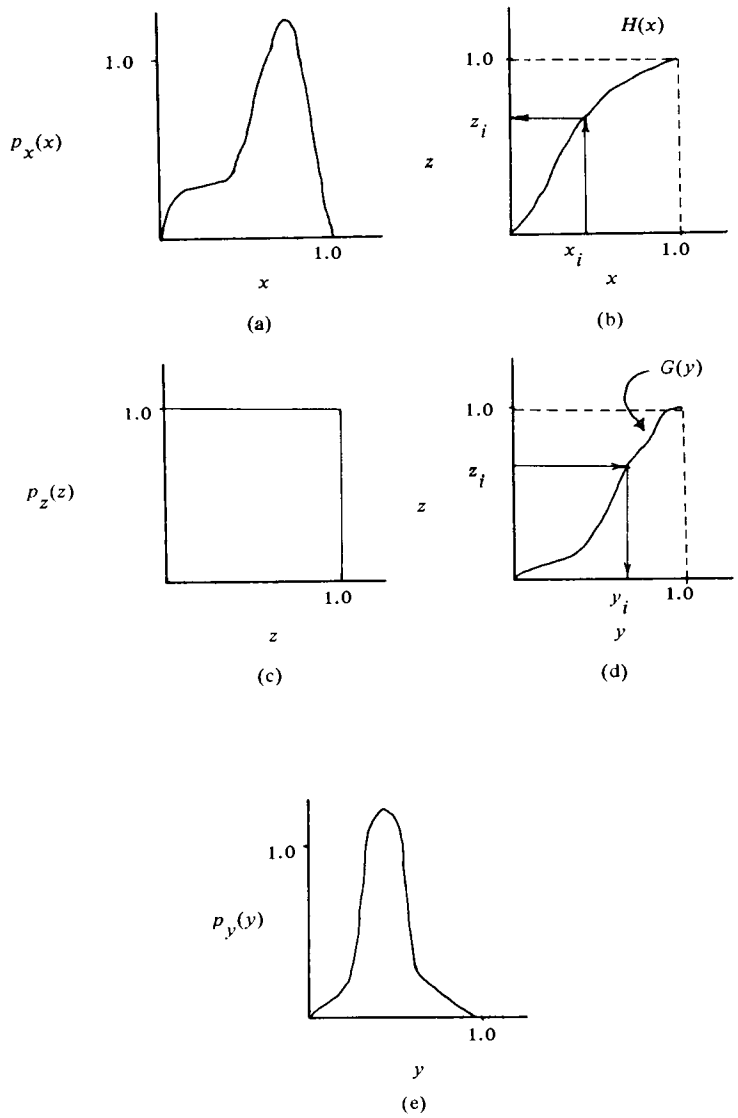
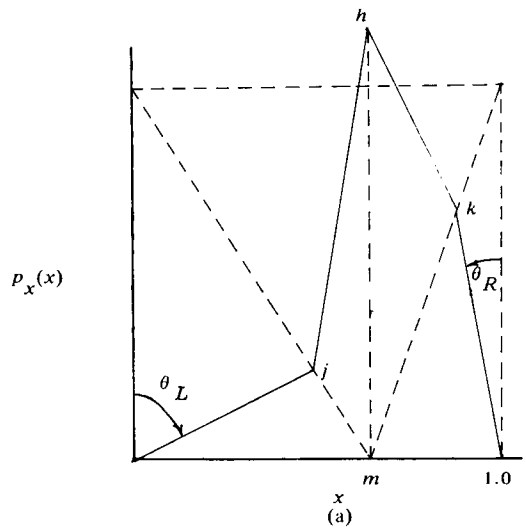
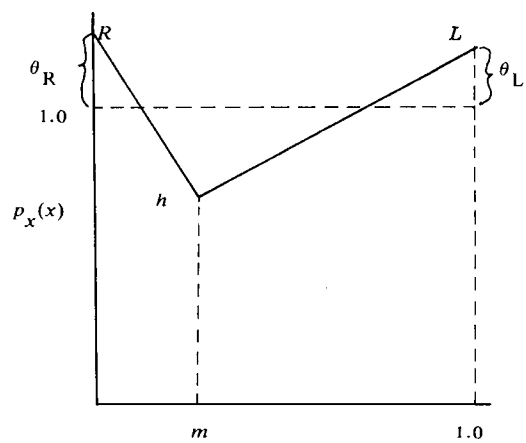


Figure 4. Histogram specification transformation procedure.



(a)



(b)

Figure 5. A four-parameter approach for specifying histograms: (a) decreasing spread. (b) increasing spread.

from 0° to 90°. Point j moves along the line segment that connects (0, 1) and (m, 0) as θ_L varies, while point k moves between (1, 1) and (m, 0) as θ_R varies. For the case of increasing spread, θ_L and θ_R specify the points along the vertical axis at which the points L and R are fixed. The lower bound for these points is zero, and theoretically there is no upper bound.

The above approach was implemented on a minicomputer, using four system potentiometers which are interfaced to the processor. Each potentiometer controls one of the four parameters described above by varying a voltage input to an analog-to-digital converter. This voltage is then sampled and expressed as a number between +1 and -1. If this algorithm were to be implemented in a real time image enhancement system, the potentiometers could be replaced by a joystick with four degrees of freedom. This would be more convenient than the potentiometers, and one could rapidly learn to manipulate the joystick in order to achieve fast enhancement results.

SPATIAL FILTERING

Spatial filtering techniques generally involve two-dimensional or regional processing of an image to produce a desired filtered effect. It is well known that lowpass spatial filtering "blurs" an image, highpass spatial filtering produces "edge enhancement" and bandpass spatial filtering "matches the spatial response of the human visual system." Of course the parameters of the filters must be correctly chosen to produce the desired effect. Spatial filtering may also be used in combination with contrast enhancement to produce excellent enhancement effects.

To introduce the problem of image enhancement by spatial filtering [3,4] we will first consider the continuous imaging situation then discretize the process for implementation in a computer. Let p(x, y) represent the original undistorted picture at the point (x, y), d(x, y) represent the observed distorted image, n(x, y) represent an additive noise process and h(x, y) represent the point spread function which characterizes the position invariant imaging system.

If there is no noise present, the imaging system will produce a distorted image, b(x, y) which may be described by the convolution equation

$$b(x, y) = \iint_{-\infty}^{\infty} h(x - x', y - y')p(x', y')dx'dy' \tag{11}$$

If additive noise is present then the distorted image is described by

$$d(x, y) = b(x, y) + n(x, y) \tag{12}$$

The continuous restoration problem may be stated as follows: given d(x, y), estimate p(x, y). The methods for solving this problem differ mainly in their assumptions and use of a priori information about h(x, y), n(x, y) and p(x, y).

To reformulate the problem for interactive computer solution two main steps are required. First, the observed image must be sampled; then the integral must be approximated by a numerical integration quadrature formula. Suppose that d(x, y) and n(x, y) are sampled at a finite set of points to produce

$$d(i, j) \text{ and } n(i, j) \text{ for } i = 1, 2, \ldots, I$$

$$u = 1, 2, \ldots, J \quad . \tag{13}$$

Sampling the noise at the same points as the observed image is simply a mathematical convenience. We may now express the distorted image as

$$d(i, j) = \iint_{-\infty}^{\infty} h(x_i - x', y_j - y')p(x', y')dx'dy' + n(i, j) \tag{14}$$

where the point (x_i, y_j) corresponds to the sample point (i, j). To complete the process, the continuous integral must be replaced by a weighted sum of the integrand values at points

$$(x'_k, y'_\ell) \text{ where } k = 1, 2, \ldots, K$$

$$\ell = 1, 2, \ldots, L \tag{15}$$

Under these conditions, the equation becomes

$$d(i, j) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} w_{k\ell}\, h(x_i - x'_k, y_j - y'_\ell)p(x'_k, y'_\ell) + n(i, j) \tag{16}$$

where $w_{k\ell}$ is the weighting factor at (x'_k, y'_ℓ). We may simplify the notation by letting

$$h(i - k, j - \ell) = h(x_i - x'_k, y_j - y'_\ell)$$

and

$$p(k, \ell) = p(x'_k, y'_\ell) \tag{17}$$

to obtain

$$d(i, j) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} w_{k\ell}\, h(i - k, j - \ell)p(k, \ell) + n(i, j) \quad . \tag{18}$$

In some cases enough information is known about the impulse response and noise process to permit an exact solution. However, in many applications it is difficult to obtain these characteristics. In these cases the interactive approach becomes very attractive.

A simple local spatial filtering may be used to realize a variety of spatial filters in a fast computation:

$$d(i, j) = \sum_{k=1}^{3} \sum_{\ell=1}^{3} h(i - k, j - \ell)p(k, \ell) \tag{19}$$

The impulse response weights, $h(i, j)$, may be selected to provide highpass filtering, lowpass filtering, high-frequency emphasis or noise smoothing. Furthermore, repeated applications of a given filter may be used to obtain more pronounced effects.

To determine if the visualization of mass objects could be recovered from an edge enhanced image by performing an inverse operation a series of experiments were made. Since edge enhancement results from highpass spatial filtering, the inverse operation is a lowpass spatial filter. Since the exact MTF was not known, a simple averaging filter was used repeatedly to obtain the desired effect. Contrast enhancement was also used. The results of these operations are shown in Figure 6. Step 1 is the original picture enhanced by contrast stretching. Step 2 is the low pass filtered image of Step 3. Step 4 is blurred twice, Step 5 is blurred three times, etc. Note that the mass lesions do indeed become pronounced by lowpass spatial filtering. The effect of spatial highpass filtering of a radiographic phantom is shown in Figure 7. The wire elements show up much clearer in the enhanced image.

PSEUDOCOLOR ENHANCEMENT

The purpose of pseudocolor enhancement is to present the original numerical image in a chromatic manner with the objective of increasing the perceptibility of important details or the speed of object recognition. The technique for mapping a single spectral representation into a three color representation is called "pseudocolor" and has been considered by many authors in several applications. The fundamental reason for considering pseudocolor is that the human visual system contains three color sensitive receptors and consequently has more capacity for distinguishing color information.

Visual chromatic stimuli must be specified by three coordinates. The red, green, and blue (R, G, B) coordinates are often used. However, for technical use, the most commonly used values are the three C.I.E.* (X, Y, Z), called "tristimulus values." These quantities may be thought of as the amounts of three standard primaries in a mixture that matches the given visual stimulus. This system of specification is so designed that one of the coordinates (Y) contains the entire brightness component of a stimulus, called "luminance," and measured in appropriate photometric units (e.g., lumens). Chromaticity can then be specified by two normalized coordinates (x, y) where:

$$x = \frac{X}{X + Y + Z} \qquad y = \frac{Y}{X + Y + Z} \tag{20}$$

The result is that all physically possible chromaticities are contained within a closed curve, known as a chromaticity diagram, on an (x, y) plot. The CIE (x, y) chromaticity diagram is shown in Figure 7, with color names assigned to various regions for ease in interpretation [8]. This curve was developed by psychovisual experiments by showing an observer a monochrome light varying from blue (450 nm) to red (700 nm) and asking him to distinguish the known from a slightly shifted unknown.

The eye can distinguish a difference in any of these three visual coordinates (2 chromaticity and 1 luminance), and is quite sensitive in this regard even for complex images. Discrimination tasks with a black-and-white image use only the eye's ability to detect differences in one of the coordinates, i.e., the luminance coordinate over its gray scale range from black to white. A pseudocolor enhancement of the black-and-white image involves the assignment of a particular chromaticity-luminance combination in the pseudocolor image to correspond to each particular shade of gray in the original black-and-white image. By using all three visual coordinates, one can transform

*Commission Internationale de l'Eclairage (International Commission on Illumination).

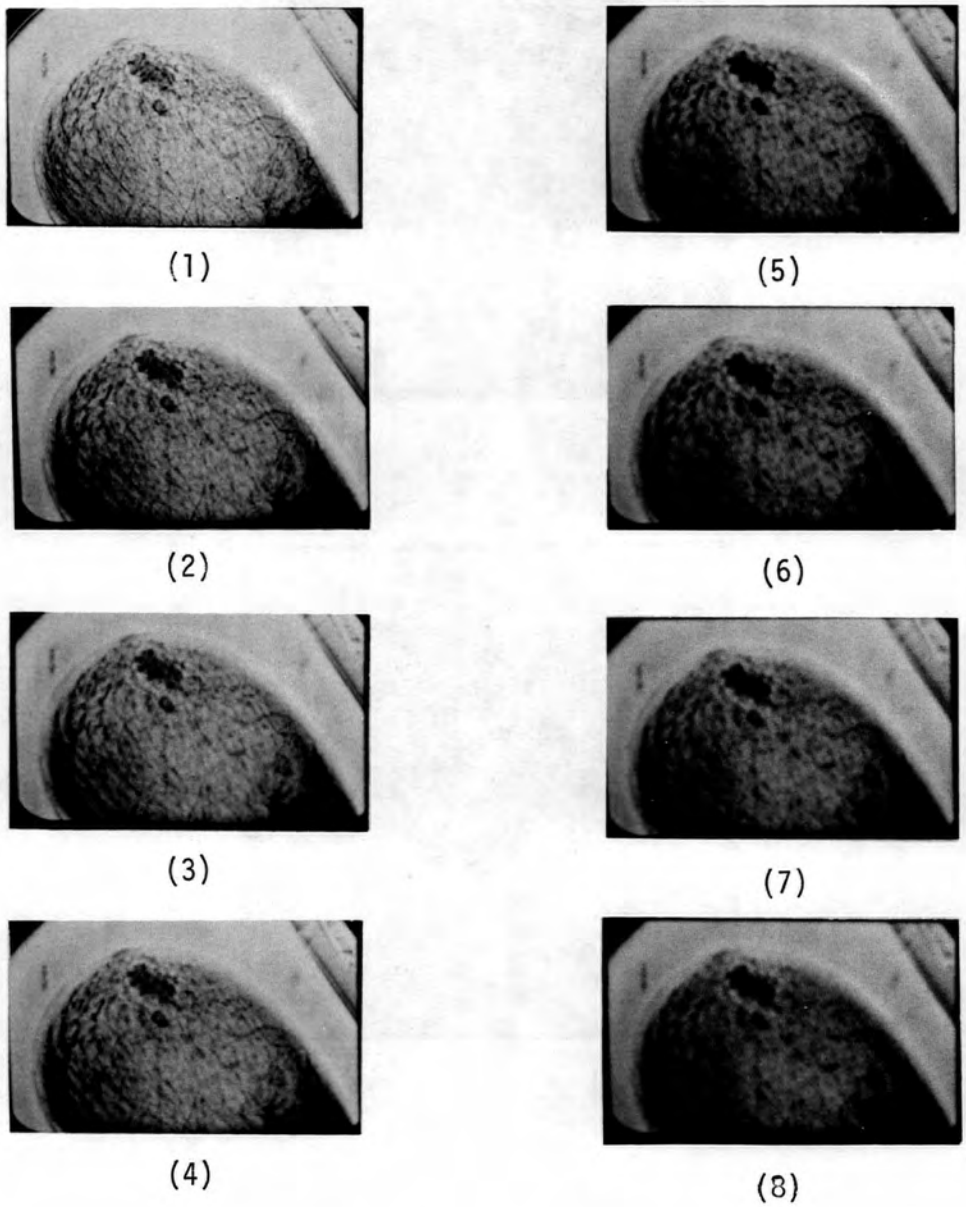


Figure 6. Enhancement of mass lesions by repeated lowpass filtering.

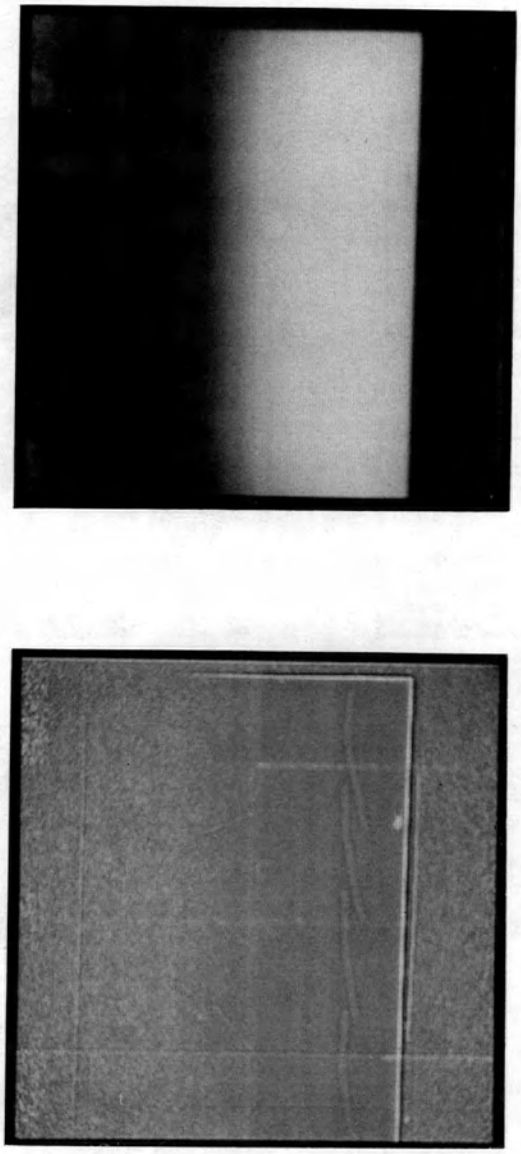


Figure 7. Effects of highpass spatial filtering.

a small gray-scale difference in the black-and-white original into an easily distinguished color difference in the pseudocolor enhancement. The specified chromaticities may be produced from the luminance values by a mapping procedure.

The pseudocolor mapping from luminosity (one dimensional space) to chromatic (three dimensional space) is difficult to define in an optimum manner not only because of the dimensionality increase but also because of the effects of color perception. For each luminosity value, Y, a set of three numbers (r, g, b) must be specified to combine the spectral responses and produce a color sensation. The arbitrariness permitted by forcing a mapping from one to three dimensions is both the beauty and the curse of pseudocolor.

Conceptually, a dimensional increase may be made in a completely specified manner if constraints are imposed on the mapping. One such constraint is a curve path. For example a line may be uniquely mapped into a spiral in three dimensions if distance along the line is made to correspond to distance along the spiral. The most familiar pseudocolor mapping occurring in nature is the rainbow. The curve path of the rainbow shown on the chromaticity diagram is similar to a parabola and is shown in Figure 8. Note that not all colors are visible in the rainbow (purple is absent) but that a large number are present. Another feature of the rainbow mapping is that since humans are familiar with it, a minimal amount of training is required to specify a scale (e.g., blue-cold; green-medium; red-hot).

Another important constraint which can be imposed on the mapping is related to perceptual distance. A perceptual distance is easily assumed for luminance perception: black is farther from white than from gray. Furthermore just perceptual differences may be determined along the range. With pseudocolor, the mapping must be defined and learned by the observer. The spectral mapping is one familiar concept. However, the just perceptual differences may be used to define a chromaticity mapping in which equal lengths along the curve correspond approximately to equal perceptual distances. This concept has lead to the CIE 1960 (u, v, w) coordinate system.

In the present state of development, the theoretical constraints on curve shape and distance are not sufficient to uniquely define an optimum pseudocolor mapping. Thus, empirical methods must be used to select an appropriate mapping for each application. Fortunately, the mappings require only a table look-up procedure for implementation and several tables may be easily stored on a computer for experimental use. In fact, each mapping can be rotated in time during display to produce an effective search method.

An experimental procedure which could test the effectiveness of pseudocolor mapping may be outlined as follows: a set of borderline images should be selected. The larger the number the more accurate the resulting statistics. These images should be mixed with normal images in a random manner. Then the complete set may be viewed by one or more observers and a receiver operating characteristic (ROC) curve developed. The ROC curves for black and white may then be compared to those obtained with pseudocolor and an objective comparison made.

The basic effects of pseudocolor can be demonstrated without an exhaustive ROC type study. The first problem encountered is selecting a mapping function. Furthermore, a particular function may be rotated in a wrap-around manner to give a large number of related functions. Starting with the computed tomography image in Figure 10(a), and using a rotated, spectral mapping show below Figure 10(b), the image shown in Figure 10(b) results. Note the enhanced visibility of the mass lesion.

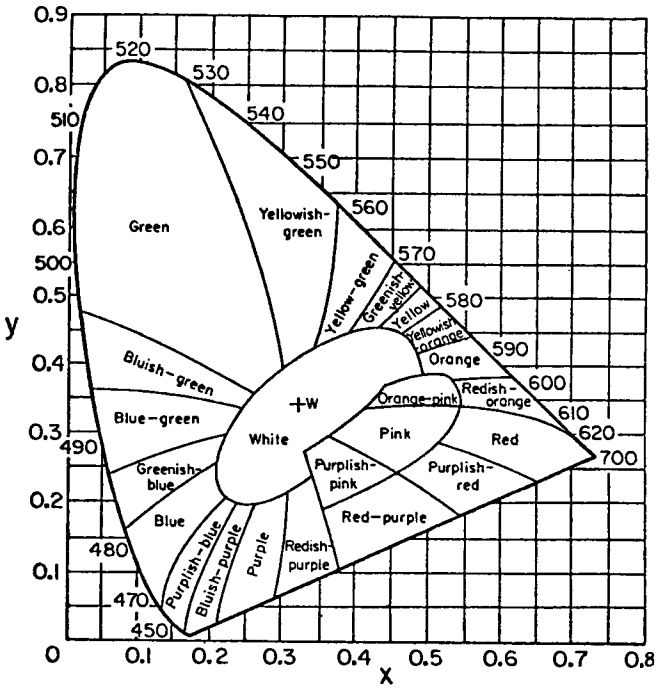


Figure 8. The C.I.E. (x,y) chromaticity diagram with color names assigned to various regions.

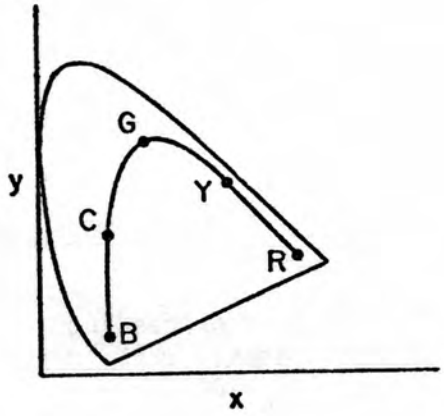


Figure 9. Curve path specified for the pseudocolor mapping.

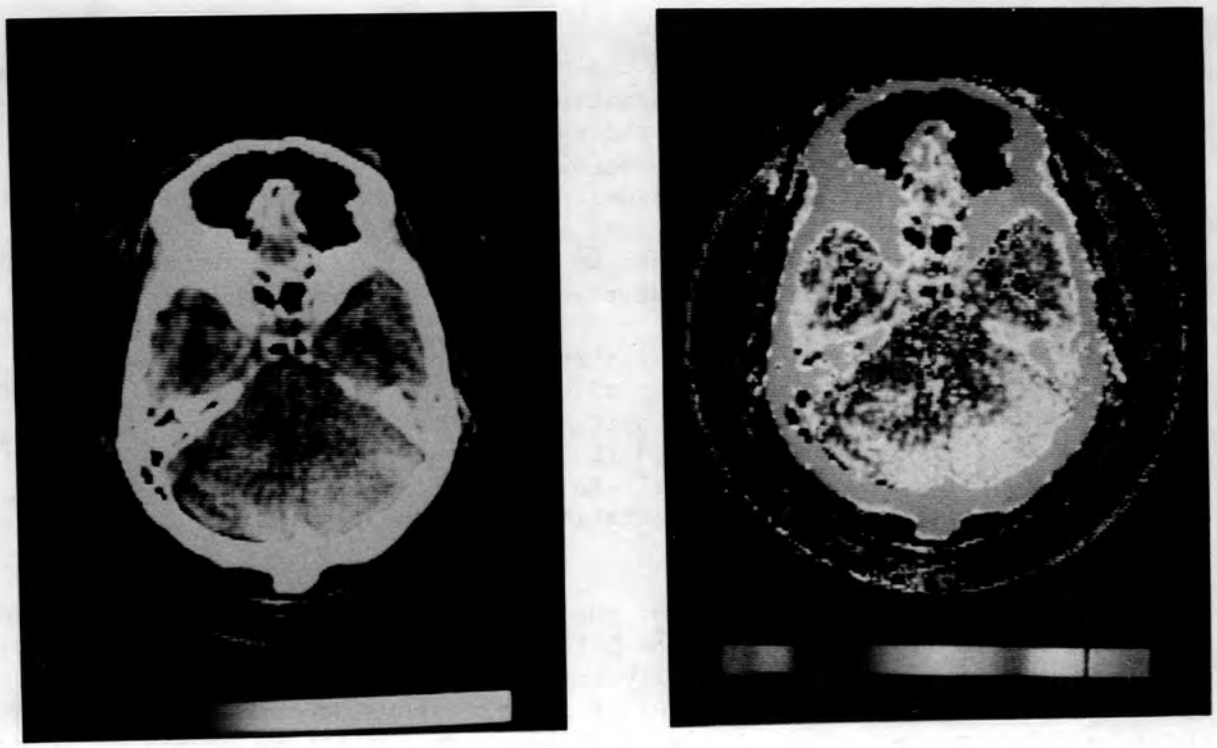


Figure 10. Enhancement of mass lesion by pseudocolor mapping.

REFERENCES

1. S. M. Pizer and R. E. Johnston, "Medical Display Systems," Technical Report, University of North Carolina, Chapel Hill, N. C.

2. R. C. Gonzalez and B. A. Fittes, "Gray-Level Transformations for Interactive Image Enhancement," Mechanism and Machine Theory, 1977, vol. 12, pp. 111-122.

3. R. C. Gonzalez and P. A. Wintz, Digital Image Processing, Addison-Wesley, Reading, Massachusetts (to appear).

4. E. L. Hall et al., A Survey of preprocessing and feature extraction techniques for radiographic images. IEEE Transactions on Computers, vol. C-20, no. 9, 1971, pp. 1032-1044.

5. E. L. Hall, "Almost uniform distributions for computer image enhancement," IEEE Transactions on Computers, vol. C-23, no. 2, pp. 207-208 (1974).

6. E. B. Troy, E. S. Deutsch and A. Rosenfeld, "Gray-level manipulation experiments for texture analysis," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 1, 91-98 (1973).

7. G. Wyszecki and W. S. Stiles, Color Science, Wiley, New York, 1967.

8. J. J. Sheppard, Jr., Human Color Perception, American Elsevier, New York, 1968.

DISCUSSION:

Kirch had the impression from the morning session that one always loses fidelity when one takes image information into the computer, and he asked how much the authors thought that the radiographs were hurt in the digitizing step. Hall agreed on this. It is only possible to handle a few line-pairs/mm whereas there are hundreds of line-pairs/mm on film. It is, however, possible to use a region selection procedure, i.e., to locate the region you are interested in and expand it in order to see greater detail. However, this type of interaction was not yet built into their system.

Kirch asked if there is any high resolution video display on which a radiologist would be satisfied to read chest x-ray film, and Hall replied that at Yale University they used for their conferences in radiology a 525-line TV system which obviously satisfied the pathologists. Brill stated that if you want to focus on a small detail and use optical enlargement and then digitize it, you preserve the information. The question is whether a radiologist would use it.

Rogers asked Hall's opinion of the most efficient way a human can interact with the computer: light-pen, track-ball, teletype or menues. Hall believed this question was unsolved generally but thought the most natural way was to point on a well-designed menue. Pizer referred to a doctoral dissertation by Edward Britton at UNC that presents a methodology for selection of interactive techniques in general. Brown commented that he was of the opinion that, for display systems, it was better to control such things as rotation, etc., by push-buttons and knobs than by menues. In this way one does not lose his concentration while looking at the image.

Todd-Pokropek commented that N. Brown showed pictures with 25 microns resolution which he considered to be less interpretable than x-rays exhibiting the same kind of resolution. This suggested that there was a loss of data with zooming after all.

Coulam responded to several of the previous points from a clinical point of view, saying that he did not know of anyone in any institute who reads radiographs from any type of televised system, whatever the resolution of the display. It is used in conferences only because this is low-level decision-making, and it is used to point out gross abnormalities and to put the whole clinical picture together. He also stated that televising an image and trying to interpret from that was ridiculous; why not look at the x-ray itself. He then emphasized the time problem, since a radiologist might read 100 x-rays a day and does not have the time to sit and zoom to all single little aspects of a film. He also expressed doubt as to whether it was economically justified to have a computer terminal at every view-box. He concluded by saying that, according to his experience, the human eye-brain combination was very good and that it is very difficult to beat it, unless you want to enhance certain aspects of an x-ray or make transformation to Fourier space, etc.

Goldberg asked what the power of the laser was in order to get the Fourier transform images and why the laser was needed at all. He stated that they have performed edge-enhancement, like band-pass filtering in Fourier domain on non-medical images,and have found that for this purpose one needs a very powerful laser; incoherent light did exactly as well. Hall replied that the power of the He-Ne laser was 15 mW and that they used the laser since they were more used to it; it was a stable light-source most of the time.

Brown concluded the discussions by stating that one thing the eye-brain complex cannot do nearly as well as the machine is to measure gray-scale density values. He thought that we should concentrate on techniques which make use of that strength of the machine.

DATA PROCESSING AND FUNCTIONAL IMAGING IN RADIONUCLIDE VENTRICULOGRAPHY *

H. Geffers, W. E. Adam, F. Bitter, H. Sigel, H. Kampmann
Universität Ulm (MNH) Department Radiologie III (Nuklearmedizin)
Ulm, West-Germany

ABSTRACT

A method is presented, which allows delineation of the end-diastolic contours using amplitude and phase of the first Fourier coefficient of local time-activity-curves. Several parameters have been used to detect local dysfunction of the left ventricular wall: phase, regional time-activity-curves, count rate differences between various frames, the correlation coefficient between local and global left ventricular time-activity-curves and maximum contraction and relaxation velocities. The validity and sensitivity of the method is demonstrated by comparison with radioopaque ventriculography and Tl-201 scintigraphy. In order to calculate left ventricular EF, two methods for evaluation of background radiation have been used. Method 1 assumes left ventricular background to be equal to the average regional background. Method 2 approximates background by an inclined plane. The influence of the selection of the background region is discussed. Comparison of the results with radioopaque ventriculography shows, that good agreement can be achieved applying the inclined plane method.

Radionuclide techniques have been widely applied since some years for assessment of global and local left ventricular performance.[1-4] In this papers emphasis will be put on the use of functional parameters to detect local wall motion abnormalities. Because left ventricular wall motion is a complex space and time dependent process, we tested several parameters on their usefulness in the detection of myocardial dysfunction. In this respect the presentation of data is very important. We found a multiple grey scale or isocontour plot very helpful, allowing a quantitative form of display. One of these parameters, the first Fourier coefficient of the local time-activity-curves is of special interest, because it allows an accurate determination of the end-diastolic heart regions.

DATA AQUISITION

Data recording begins several minutes after intravenous administration of about 20 mCi of Technetium labeled HSA. Fifteen 32 x 32 matrices are produced in the computer memory, which, beginning with the R-wave of the ECG, represent

* Supported by the German Ministry of Research and Technology (BMFT)

different time intervals covering the whole heart cycle. 350-600 beats are recorded omitting post-extrasystolic cycles. Variation of the RR-time is in a global way taken account of by renormalizing each matrix in order to have the same number of recorded cycles for each time interval. We use a high sensitivity collimator in order to reduce data collection time.

DATA ANALYSIS

Determination of the Heart Regions

The contours of the heart cavities can easily by determined by taking advantage of the properties of the first Fourier coefficient. It is given by the formula:

$$F_1 = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} A_n \, e^{-\frac{2\pi i n}{N}} ,$$

$$= X + i\,Y \quad .$$

N is number of time intervals, the A_n are count rates at time interval n. Assuming the count rates to be Poisson distributed, the components X,Y of F_1 satisfy to a good approximation a shifted Gaussian distribution.

$$W\,(X,Y) = \frac{1}{2\pi s^2}\, e^{-\frac{(X-X_o)^2 + (Y-Y_o)^2}{2s^2}} ,$$

$$s^2 = \frac{1}{N} \sum_{n=1}^{N} A_n \quad .$$

The statistical standard deviation s is given by the square root of the mean local count rates. X_o, Y_o are the expectation values of X and Y. They are zero if there is no motion at the respective image point. The expectation value of the amplitude r is given by

$$r^2 = X_o^2 + Y_o^2 + s^2 = r_o^2 + s^2 \quad .$$

Moving structures can easily be discriminated against not moving structures pres cribing a lower threshold greater than s for the amplitude. Thus cutting off statistical noise of not moving structures. Fig. 1 shows on the left side the amplitudes for a normal subject from LAO 30° projection. While all other structures like liver and lung reduce to the flat noise level, the heart cavities, left and right ventricle and the atria together with a part of the large vessels, emerge very clearly. The phase distribution is uniform if r_o is zero (white noise) and becomes very sharp with increasing r_o. For $r_o > s$ the standard deviation of the phase is approximately given by:

$$\overline{(\varphi - \varphi_o)^2} = \frac{s}{r_o} \qquad \text{for } r_o > s \quad .$$

These properties of the phase are seen on the right side of figure 1. One sees the ventricles as a nearly flat region of constant phase, the atria and large vessels, whose phase differs by about π from the phase of the ventricles, and

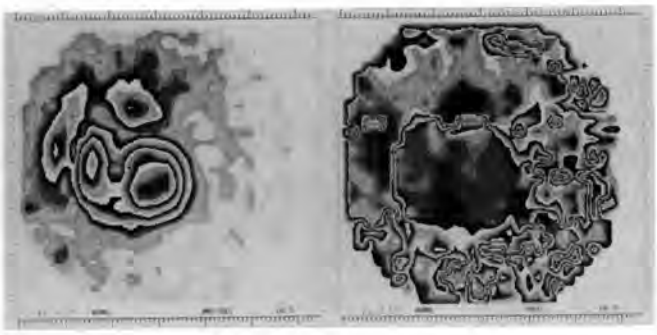


Fig.1. Fourier amplitudes (left) and phases (right) of a normal subject from LAO 30° projection. Black corresponds to 1/4, 1/2 and 3/4 of the maximum amplitude resp. to $\pi/2$, π, $3\pi/2$ and 2π.

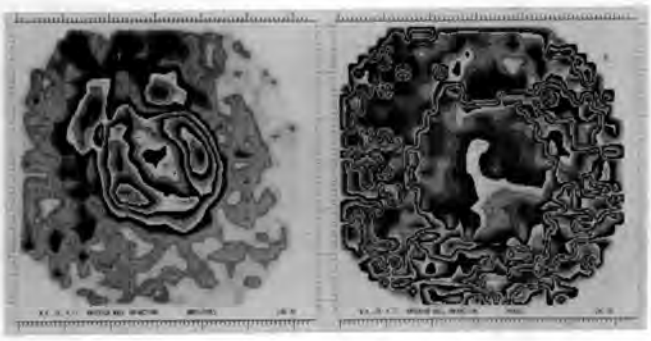


Fig.2. Fourier amplitudes (left) and phases (right) of a patient with an old anterior wall infarction and an aneurysm at the apex.

the background region with its chaotic distribution of the phase. By using amplitude and phase together the enddiastolic region of the ventricles may be determined to a high degree of accuracy, if simultaneously a lower threshold for the amplitude and a small interval for the phase is prescribed. The adequate phase interval is determined by taking the phase of the image point with maximum amplitude as the center and adjusting the width of the interval. The method works equally well for normokinetic subjects as for patients with severe reduction of left ventricular performance. In order to complete the contour if akinetic scars occur at the border of the left ventricle, the begin of the steep ascent of enddiastolic count rates is determined. Fig.2 shows amplitudes and phases for a patient with an aneurysm and a large akinetic scar of the anterior wall. The method relies mainly on statistical reasons and is therefore principally independent of the observer, it is simple and may be applied to arbitrary projections. The method can be extended to a fully automatic determination of the enddiastolic heart regions.

Local Functional Parameters

Correlation Coefficient and Phase

We spent a large part of our work in searching for parameters by which local dysfunction of wall motion can be detected to a high degree of sensitivity because we are interested in the question if there exist early signs of altered contraction behaviour in patients suffering from coronary heart disease. Such an attempt is reasonable only if the local time-activity-curves are given with a sufficient accuracy. Investigation of more than 20 normokinetic subjects showed us that the correlation between local and global time-activity-curves is extremely high, provided that the mean count rate in the left ventricular region is of the order of 600 or higher. Fig.3 shows on the left side the correlation coefficients for a normal subject. All image points except for some points at the basis or the outer border have correlations greater than 0.9 and 85 percent of the inner points have even correlations greater than 0.95 in this case. This shows that time-activity-curves from single image points reflect very well the synchronous motion of the left ventricle and suggests that even small local deviations within the behaviour of time-activity-curves may have statistical significance. Within infarcted regions the correlation coefficients may drop to low values of the order of zero and even may become negative in aneurysms (Fig.3).

Thus the correlation coefficient can be used as a functional parameter, but in practice it is not very adequate, because its statistical properties are not

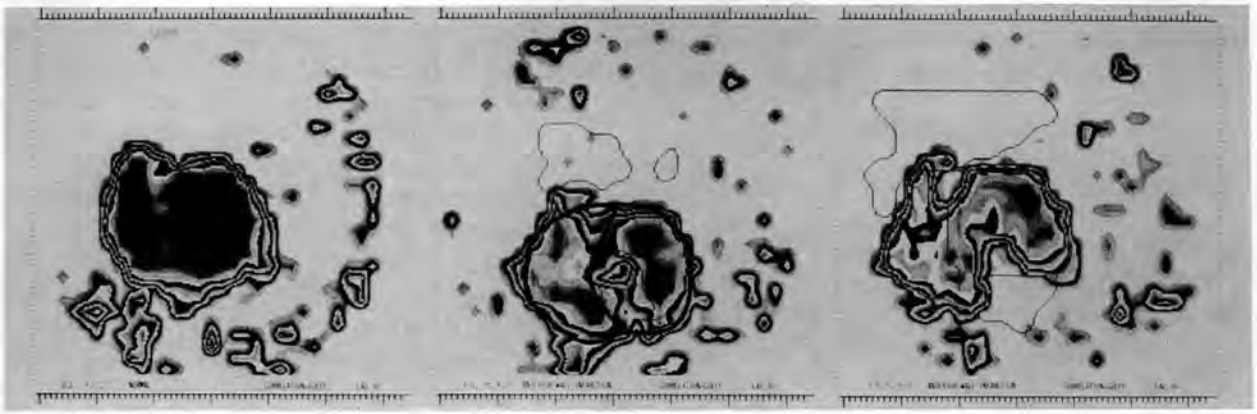


Fig.3 Correlation coefficients between local and global left ventricular time-activity-curves. Only positive correlations are shown. Black corresponds to 0.25, 0.50, 0.75 and 1. Normal subject at left, fresh anterior wall infarction (middle), aneurysm with negative correlation (right).

easy to handle. A much better parameter in this respect is the phase of the first Fourier coefficient. For normal subjects, as we have seen before, the phase is nearly constant within the left ventricle except for small statistical fluctuations and often a small drift from the basis to the apex. Connected regions of deviating phase reflect contraction abnormalities. It is very useful to mark off such regions for calculation of time-activity-curves. A simple statistical evaluation can be made in order to check if there is a significant deviation between the regional and the global curve. The vector component of the first Fourier coefficient of a regional curve, which is orthogonal to the first Fourier coefficient of the global curve, should be Gaussian distributed with zero mean if both curves have the same time behaviour. If that is not the case both curves are significantly different from each other. In order to demonstrate that very small differences in time behaviour can be detected by this method, Fig.4 shows amplitudes and phases of a patient suffering from aortic valve stenosis and regurgitation. At the apex there is a region of slightly reduced amplitudes and altered phases. The time-activity-curve shows that contraction begins later in this region suggesting myocardial damage due to permanent pressure overload of the left ventricle. These findings could be confirmed by cineangiography. Similar results but without reduction of the amplitude did sometimes occur for patients with mitral valve stenosis.

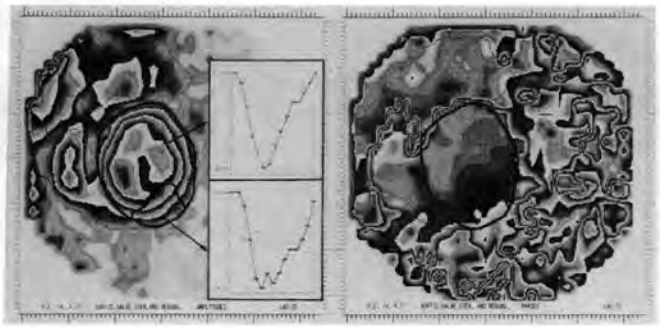


Fig.4 Fourier amplitudes and phases of a patient suffering from aortic valve stenosis and regurgitation. Regional curves have been calculated from the marked off region.

Count Rate Differences

The most direct parameters to asses regional dysfunction are count rate differences. Count rate differences between the enddiastolic and the endsystolic frame for example are equivalents of the exent of systolic contraction. We plot these differences in such a way that one or two statistical standard deviations, approximately given by square root of twice the average left ventricular count rate, correspond to one grey scale in our multiple grey scale plot. By this way unwanted statistical noise can be suppressed and the risk of false positive results be reduced, while the important information is preserved. Fig.5 shows systolic contraction for a normal subject, a patient suffe ring from aortic valve stenosis and two patients with old anterior respectively posterior wall infarction. Normokinetic subjects show a typical concentric pattern from LAO 30° projection. Slight irregularities in normal subjects do sometimes occur in the aortic valve region. Scars and regional hypokinesy can easily be detected by an irregular contraction pattern. Sometimes abnormalities are more distinct in diastolic relaxation than in systolic contraction. Therefore we plotted count rate differences between the frame, which corresponds to the beginning of the diastasis, and the end-systolic frame. The diastasis is the flat portion of the left ventricular time-activity-curve. For normal subjects the relaxation pattern was quite similar to the contraction pattern. For patients with severe local dysfunction contraction and relaxation pattern were abnormal, but there was a small group of patients which had only an abnormal relaxation pattern.

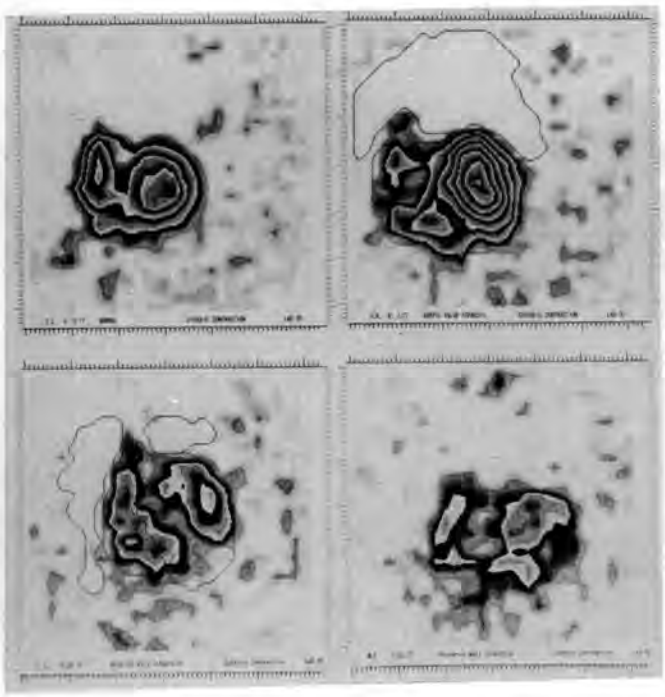


Fig.5 Extent of systolic contraction of a normal subject (top left), a patient suffering from aortic valve stenosis (top right), a patient with an anterior wall infarction (bottom left) and a patient with posterior infarction (bottom right).

Contraction and Relaxation Velocities

Because radionuclide ventriculography yields time-volume-curves to a degree of accuracy not attained by other methods, one may hope that analysis of these curves will yield further valuable clinical parameters. There exist strong hints that a reduced velocity of fiber lengthening during early diastole occurs in many patients with coronary artery disease or with mitral valve stenosis. [5] Fig.6 shows a typical left ventricular curve for a normal patient and a curve for a patient with coronary artery disease. The slope of the diastolic part of the curve is strongly reduced, perhaps due to microscopic necrosis and stiffening of the myocardium. In order to investigate the regional pattern of relaxation and contraction velocities, we calculated time derivatives of the local curves after applying a Wiener filter. The derivatives have been taken at the inflection points of the global curve. For normal subjects there is a common pattern similar to that of systolic contraction (Fig.7). If local dysfunction

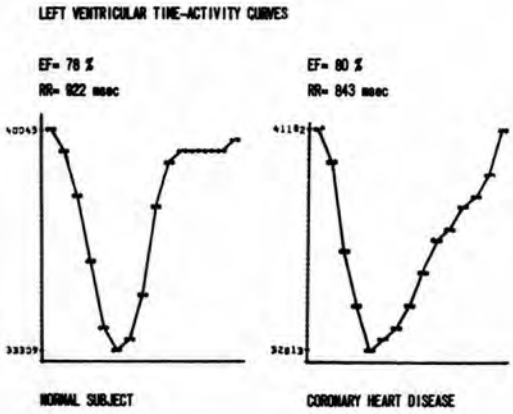


Fig. 6. Left Ventricular Time-Activity Curves.

can be detected by other parameters, the local contraction and relaxation velocities show analogous defects as is to be expected. Fig.8 shows as an example a patient with an old intramural supraapical infarction. In some cases defects were more pronounced for these parameters than for the extent of contraction or relaxation. Patients suffering from coronary heart disease which are normal with respect to the other parameters show no drastic deviation from the normal pattern. A final conclusion about the clinical usefulness of these parameters is not possible at present.

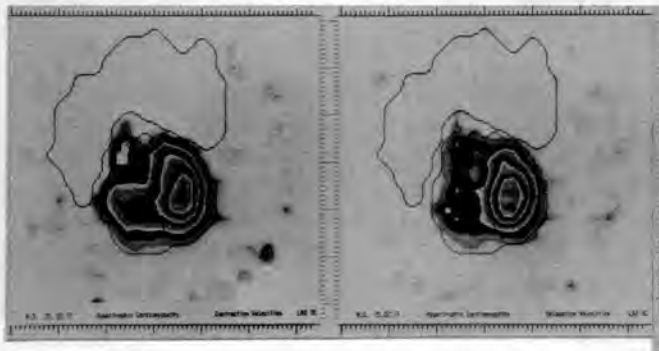


Fig.7 Contraction velocities (left) and relaxation velocities of a normokinetic patient (hypertrophic cardiomyopathy).

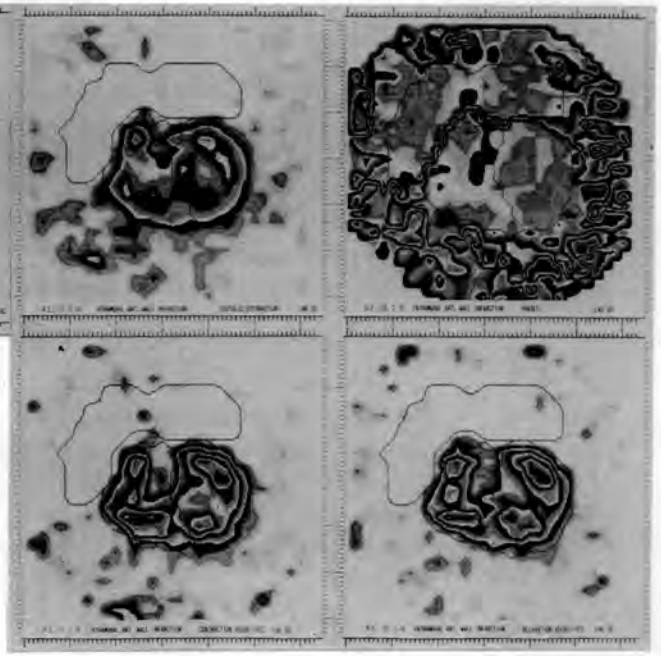


Fig.8 (Right) Systolic contraction (top left), phases (top right), contraction velocities (bottom left) and relaxation velocities (bottom right) of a patient with an old intramural supraapical infarction.

COMPARISON OF RADIONUCLIDE VENTRICULOGRAPHY WITH RADIOOPAQUE ANGIOGRAPHY AND MYOCARDIAL SCINTIGRAPHY

The previously mentioned parameters served us to evaluate local contraction behaviour. To check the validity of the method comparison has been made for 54 patients with X-ray ventriculography. The results given in table 1 show good agreement. No false positive results have been obtained, three small scars of the posterior wall and one scar at the apex escaped detection. In some cases disagreement existed about the question, if a scar should be described as dyskinetic or akinetic, because time-activity-curves showed paradoxical motion during the early systole, wheras in X-ray ventriculography the scar seemed to be akinetic. Included in the study were only LAO 30° radionuclide ventriculographies. X-ray studies were either monoplane RAO or biplane RAO and LAO. Because of the

Table 1.
Comparison between radionuclide and radioopaque ventriculography (54 patients).

Angiography	Local Dysfunction	Local Dysfunction	Normal	Normal
Scintigraphy	Local Dysfunction	Normal	Local Dysfunction	Normal
Number of Cases	27	4	0	23

quasi three dimensional representation of the data, large posterior wall defects are visible from LAO 30°. For an exact localization and for detection of small posterior wall defects at least one additional projection is necessary.

Further on we made a comparison between radionuclide ventriculography and TL 201 myocardial perfusion studies. 65 patients underwent myocardial perfusion study with TL 201 in the resting state and radionuclide ventriculographie at the same day. All cases were proven myocardial infarctions. 28 patients had suffered acute myocardial infarction from 2 days up to six weeks prior to the investigation, 37 patients had an old infarction. Table 2 shows the results. In 49 cases perfusion defects and contraction abnormalities were present. 11 cases

Table 2.
Comparison between radionuclide ventriculography
and TL-201 myocardial scintigraphy for 64 patients with proven infarction

Tl^{201} Scintigram	Perfusion Defect	Perfusion Defect	No Defect	No Defect
Radionuclide Ventriculography	Local Dysfunction	Normal	Local Dysfunction	Normal
Number of Cases	49	1	11	4

showed significant local dysfunction, in most cases hypokinesy, whereas no sure perfusion defect could be detected, 4 cases showed neither perfusion defect nor local dysfunction and only one case showed a perfusion defect at the posterior wall, whereas no contraction abnormality could be detected. The sensitivity of TL-201 myocardial scintigraphy with respect to the detection of myocardial infarction, which we found, agrees well results reported from others.[6] The high sensitivity of radionuclide ventriculography is a strong argument for its application in the diagnosis of acute and chronic infarction.

The calculation of the left ventricular EF from LAO projection assumes an accurate evaluation of tissue background. Because tissue radiation cannot be assumed to be homogeneous, results depend on how the background region is selected. In order to investigate this influence we have chosen three different regions.

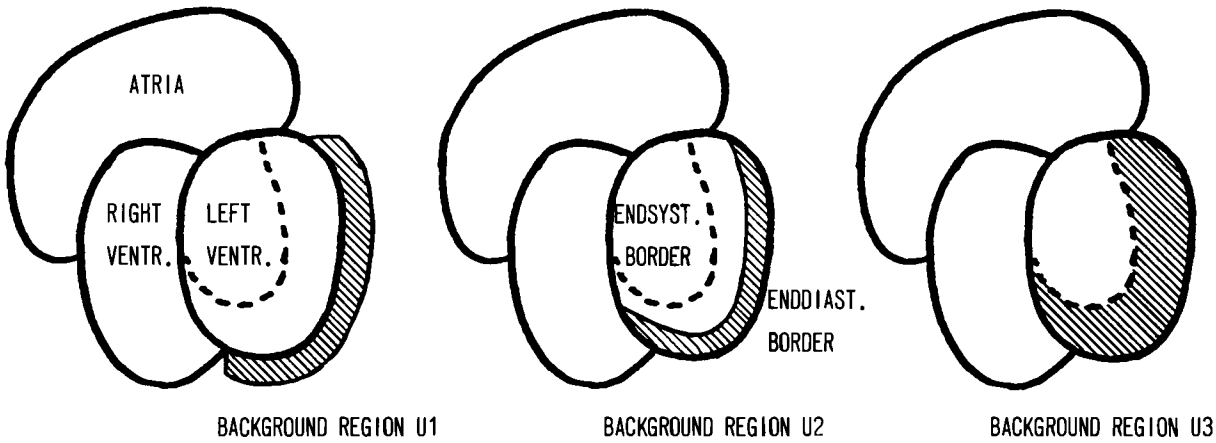
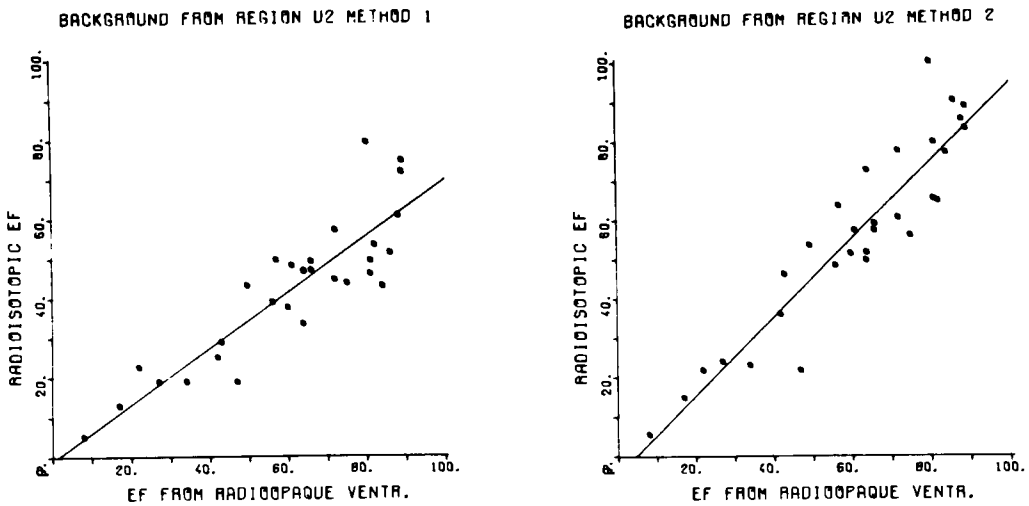


Fig.9. Background regions used for calculation of the EF.

U 1: A narrow region extending from the basis to the inferior septum just outside the enddiastolic contour.
U 2: A narrow region extending from the basis to the inferior septum just inside the enddiastolic contour in normal contracting parts or just outside in akinetic parts or dyskinetic parts of the left ventricle.
U 3: The region between endsystolic and enddiastolic border of the left ventricle. The endsystolic border has been determined as the begin of the steep ascent of endsystolic count rates.

We applied two different methods. The first method consists in taking the background to be equal to the average count rate in the selected region. This method has been widely used by other authors too.[7,8] The most simple step beyond this assumption is to approximate endsystolic count rates within the selected region by an inclined plane. The parameters of the plane have been determined by a least square fit procedure. Extrapolation of the plane to the whole ventricle yields the second approximation for background radiation. Before calculating the EF, the distances of endsystolic count rates from the plane expressed in multiples of the statistical standard deviation are displayed. By this way image points within the liver, the spleen or the endsystolic ventricle can be excluded. This procedure is necessary because of the strong influence of such points on the results. In this way for each patient six values of the EF have been obtained. A comparison is made for 30 patients with results from radioopaque ventriculography (area-length method). Fig.10 shows the results for the region U2 and for both methods. The results of the regression analysis and the correlation coefficients for all the values are given in table 3. For all selected regions the slope b of the regression line is significantly lower than 1 for the method 1 (left side of table 3). The values of b increase slightly if the selected region approaches the endsystolic contour. This supports the assumption of method 2 that background is inhomogeneous. The results of the second method are summarized on the right hand side of table 3. In this case the values are nearly centered around the line of indentiy. There is no significant

difference among the results from the three regions. The correlation coefficients in this method are slightly higher than those of method 1. These results show that the left ventricular EF can be obtained to a sufficient accuracy and in good agreement with X-ray ventriculography by the inclined plane method. This method requires a careful selection of the background region if there is much activity within the spleen and if the spleen is near the left ventricular apex.



$EF_{Scint.} = a + b * EF_{Ang.}$

a = - 0.1

b = 0.71

r = 0.886

N = 30

a = - 4.0

b = 0.99

r = 0.924

Fig.10. Scintigraphic EF versus EF angiographic EF.

Table 3.

Background Evaluation by

	Average			Inclined Plane		
Region	U1	U2	U3	U1	U2	U3
a	- 0.01 ± 0.05	- 0.01 ± 0.05	- 0.02 ± 0.05	- 0.07 ± 0.06	- 0.04 ± 0.05	- 0.03 ± 0.06
b	0.68 ± 0.07	0.71 ± 0.07	0.74 ± 0.08	1.03 ± 0.09	0.99 ± 0.08	0.99 ± 0.09
r	0.876	0.886	0.879	0.900	0.924	0.895

REFERENCES

1. G. Hoffmann, N. Kleine, Eine neue Methode zur unblutigen Messung des Schlagvolumens am Menschen über viele Tage mit Hilfe von radioaktiven Isotopen, Verh. Dtsch. Ges. Kreislauff. 31, 93 (1965).

2. B. L. Zaret, H. W. Strauss et al., A Noninvasive Scintigraphic Method for Detecting Regional Ventricular Dysfunction in Man, N. Engl. J. Med. 284, 1165 (1971).

3. D. Van Dyke, H.O. Anger et al., Cardiac Evaluation from Radioisotope Dynamics, J. Nucl. Med. 13, 585 (1972).

4. W. E. Adam, G. Meyer et al., Camera-Kinematography: a Nuclear Medicine Procedure for Imaging Heart Kinetics, J. Nucl. Biol. Med. 18, 59 (1974).

5. K. E. Hammermeister, J. R. Warbasse, The Rate of Change of left Ventricular Volume in Man, Circ. 49, 739 (1974).

6. F. J. Th. Wackers, J. B. van den Schoot et al., Value and Limitation of Tl-201 Scintigraphy in the Acute Phase of Myocardial Infarction, N. Engl. J. Med. 295, 1 (1976).

7. R. H. Secker-Walker, L. Resnick et al., Measurement of Left Ventricular Ejection Fraction, J. Nucl. Med. 14, 798 (1973).

8. M. V. Green, W. R. Brodley et al., Count Rate Measurement of Left Ventricular Ejection Fraction from Gated Scintigraphic Images, J. Nucl. Med. 17, 557 (1976).

DISCUSSION:

The paper presented by Horst Geffers demonstrated several innovative concepts for developing functional images of the left ventricle from gated scintigraphic studies. This approach has been designed to exploit parameters which are sensitive to changes in the contractional behavior of the left ventricle on a regional basis. One approach to this problem involved images based on the fundamental Fourier coefficients at the frequency corresponding to the heart rate of local time-activity curves developed for each image element. By this technique the localized Fourier coefficients computed for each image pixel are compared to a global coefficient developed from the time-activity curve for the total left ventricle. In this manner functional images are developed comparing the phase of the fundamental Fourier coefficients on a regional basis which are very sensitive to regions which lag or counter pulsate with respect to the global ventricular behavior. Functional images can also be developed on a regional basis which compare the fundamental Fourier coefficient amplitudes. Another type of regional functional image was developed by computing correlation coefficients for each image element with respect to the global left ventricular time-activity curve. A related study evaluated background estimation techniques for left ventricular ejection fraction measurements. This work compared estimates based on an annular region-of-interest surrounding the left ventricle with a background determined from a biased plane over the LV region-of-interest. The biased plane technique resulted in a significant improvement in the correlation between scintigraphic and contrast ventriculographic methods for determination of ejection fraction.

In the discussion that followed this paper, concern was expressed about the validity of applying two-dimensional Fourier analysis to the motion of a three dimensional object (the left ventricle). That is, would the method be as sensitive to regional wall motion abnormalities in the center of the ventricular image as it would be to abnormalities along the edge of the ventricular outline? It was generally acknowledged that, although the contractility maps would definitely change as a function of the projection used, the detection of non-moving or counter-pulsating regions in the center of the ventricular projection would be possible by these methods. This potential gives the scintigraphic method an added dimension of detectibility which is not achieved in methods which look at the left ventricular outline alone. It was also stated that higher order Fourier coefficients had not been evaluated as a basis for developing this type of functional maps. The use of the biased plane to develop LV background was acknowledged as a significant step in reducing the subjectivity of scintigraphic ejection fraction determination in addition to improving the correlation with contrast ventriculography.

A TECHNIQUE FOR ITERATIVE ESTIMATION OF LEFT HEART EJECTION FRACTIONS AND REGURGITANT FRACTIONS FROM NONINVASIVE RADIOCARDIOGRAPHIC DATA

D. B. Twieg
University of Texas Health Science Center at Dallas
Dallas, Texas, U.S.A.

ABSTRACT

Scintigraphic equilibrated cardiac blood pool data and radiocardiographic (first-pass) data contain information of differing types. When analyzed in terms of cardiac pump function, equilibrated blood pool data can give information about total left ventricular ejection fraction, but not about forward and regurgitant flows across the mitral and aortic valves. However, radiocardiographic data can provide information of both types. The latter fact has been demonstrated in studies by Kirch, Metz and Steele in which z-transform analysis of first-pass data obtained following injection of a tracer bolus via pulmonary wedge catheter yielded left heart forward and regurgitant fractions.

A somewhat similar technique for analyzing radiocardiographic data obtained noninvasively, with peripheral intravenous injection, might be of value in such applications as repeated monitoring of patients with valvular disease, and in investigating the temporary mitral regurgitation which occurs in the course of some acute myocardial infarctions. For this purpose we have developed a simple two-phase discrete-time model of the left heart which permits an arbitrarily-shaped tracer input function. The model is used in conjunction with a conjugate-gradient parameter-estimation algorithm, which estimates four parameters of cardiac pump function: forward ventricular ejection fraction, total ejection fraction, mitral regurgitant fraction, and ratio of maximum ventricular volume to maximum atrial volume. Form the first three of these, aortic regurgitant fraction may be computed.

Briefly, the technique involves acquisition of first-pass list mode data for forty-five seconds following peripheral intravenous injection, with the gamma camera placed precordially in the 40° LAO position with 15° caudal modification. Forty-five serial sets of end-diastolic and end-systolic activities are then computed from the list-mode data for four areas of interest: lung, left atrium, left ventricle, and ventricular background. After subtracting background and temporally shifting the input curve generated from lung activity, the iterative conjugate gradient search algorithm is applied to minimize a weighted least-squares cost function.

In order to investigate the ability of the method to estimate forward and regurgitant fractions from data obtained with a broad input function as would occur after peripheral injection, simulation studies

were performed. Noisy simulated data corresponding to the responses of six systems with different sets of parameters with bolus and broad inputs, were generated on a small computer (DEC PDP-8). The parameter estimation algorithm was then applied to the simulated data, and in all cases, even the broad input functions, the actual paramaeter values were closely approximated by the estimates.

At present, preliminary work is being undertaken to apply the technique to patient data.

INTRODUCTION

A new method will be presented for estimating several parameters of cardiac function, including regurgitant fractions, from radiocardiographic data. The significance of this method is that it is designed to operate on data acquired from noninvasive studies, which are safer and less expensive than other studies yielding the same information. This method can be used in some clinical situations where other methods cannot, such as in the immediate post-infarction period, when the heart may be very sensitive to catheterization of the left ventricle.

A BRIEF OVERVIEW OF RADIOCARDIOGRAPHIC AND RADIONUCLIDE ANGIOGRAPHIC DATA ANALYSIS

Let us consider the process of obtaining radiocardiographic or radionuclide antiographic data in the form of serial gamma camera images of tracer radioactivity in the cardiac blood pool and then applying some analytic technique to extract clinically useful information from the data. The relationships of clinically useful information to radiocardiographic and gated blood pool data and the analysis of this data is summarized in Figure 1 for noninvasive studies.

First-pass data is acquired in the form of sequential images from a precordially positioned gamma camera as the tracer passes through the heart for the first time. Gated blood pool data is acquired after the vascular tracer has achieved constant concentration throughout the blood and the gamma camera is turned on and off (gated) in a preselected time sequence according to the patient's electrocardiogram. There is a fundamental difference between the two types of data in terms of their information content, and this difference has important implications in the determination of transvalvular flows.

Equilibrated blood pool data consists of a set of two or more sequential images, each corresponding to a different time frame of the cardiac cycle. Sequential frames are recorded during each cardiac cycle, then each one is summed with all others in subsequent cardiac cycles having the same time frame or phase. The information to be extracted from this data may be referred to as "high-frequency" information, since it reflects the cyclic changes in the volumes and shapes of the cardiac chambers. Some of this information pertains to regional myocardial function, and some of it relates to total ventricular performance. We can express the regional information in terms of segmental wall motion measurements and the information on total performance, with which we are concerned here, as ejection fraction (E.F.), ejection rate, and end-diastolic and end-systolic volumes. Ventricular volume at a given phase of the cardiac cycle is often measured from high-frequency data by the geometric method; that is, by estimating the outline of the left ventricle, then applying a geometric model of the ventricle

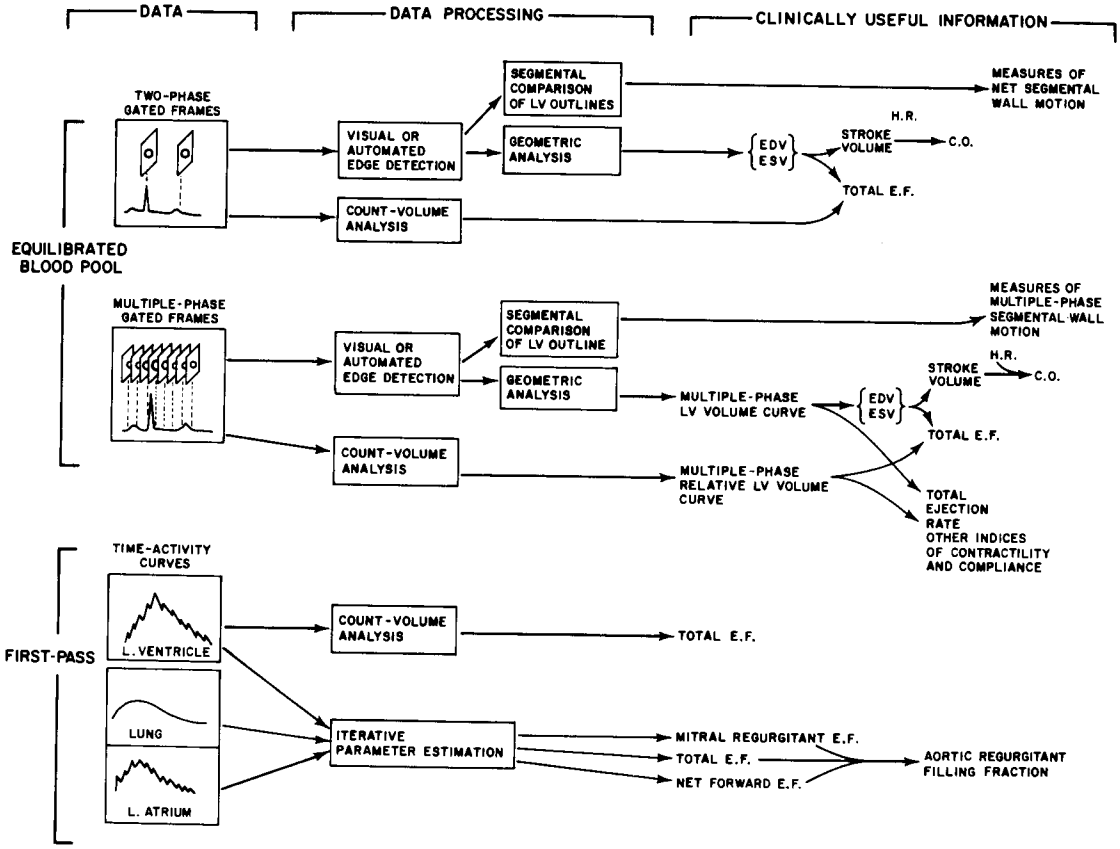


Figure 1

Methods for obtaining clinically useful information regarding myocardial function and cardiac pump function from noninvasive first-pass and equilibrated blood pool data.

to its two-dimensional image (1,2,3,4). Stroke volume and ejection fraction may then be computed by comparing the resulting estimates of end-diastolic volume (EDV) and end-systolic volume (ESV):

$$EF = \frac{EDV-ESV}{EDV} \qquad (1)$$

When the heart is viewed from an angle such that the left ventricle is isolated from other blood-containing structures (modified left anterior oblique) (5,6,7), ejection fraction may be computed from equilibrated blood pool data by comparing end-diastolic left- ventricular activity and end-systolic activity (8). This is possible because under these conditions, left-ventricular activity is proportional by a constant coefficient to left-ventricular volume.

First-pass data consists of high-temporal-resolution images of the cardiac blood pool, obtained while the tracer first passes through the left ventricle in a relatively short period of time. This series of images is usually converted to the form of "time-activity curves"*, which represents the history of activity in designated areas of the image corresponding to cardiovascular structures such as left ventricle, right ventricle, left atrium, lung, etc.

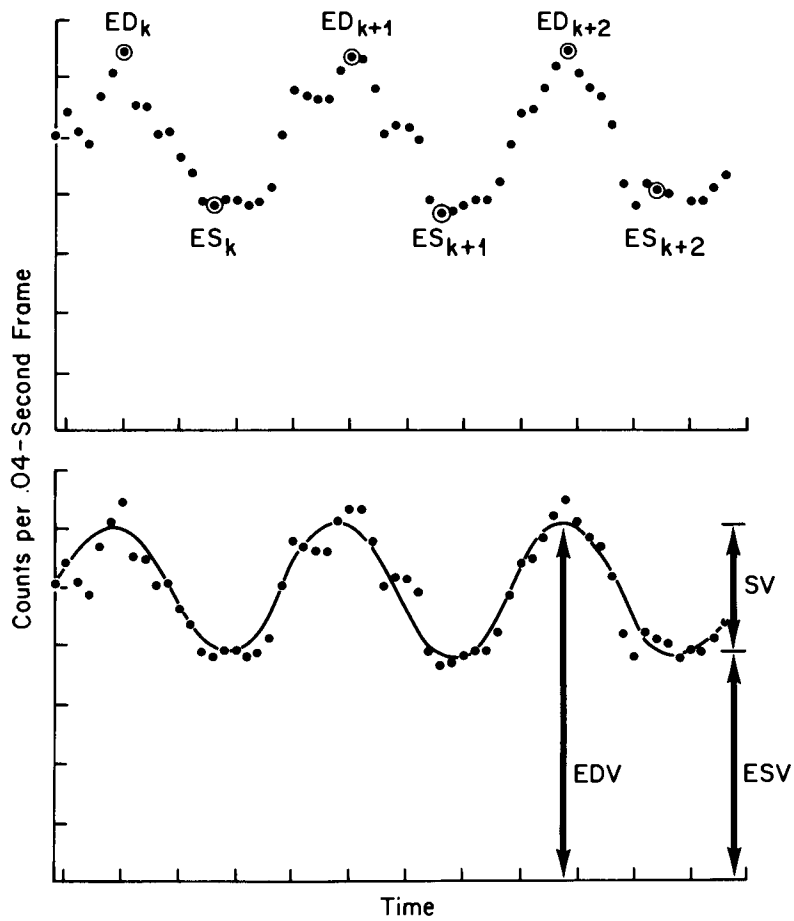
The time-activity curves contain both high- and low-frequency information, the latter representing the characteristics of the heart as an ideal mixing chamber (12) or as a pump. One consequence of both types of information being repsent is that forward and reverse flows between chambers can be determined from the data. The situation is directly analogous to that in cardiac contrast angiography, where forward cardiac output from dye-dilution curves (low-frequency information) must be compared with total cardiac output from ventriculographic stroke volumes (high-frequency information) in order to estimate valvular regurgitation (13).

CURRENT MEASUREMENT OF EJECTION FRACTION AND VALVULAR INCOMPETENCE

Total left ventricular ejection fraction (E.F) is frequently used as an important index of cardiac function. Classically, E.F. has been determined by contrast angiography alone, though radiocardiography and echocardiography are now being used (2,6,7,14,15,16,17,18). The basic premise of first-pass E.F. analysis is that under the conditions of constant ventricular tracer concentration which apply during the systolic portion of the cardiac cycle when the tracer is well-mixed, the left-ventricular tracer activity is proportional to its volume. Thus, a comparison of relative ventricular activity during consecutive end-diastolic and end-systolic time frames can be used to determine total ejection fraction. This is the case regardless of whether or not the input and output valves are competent.

Ventricular tracer activity is measured either directly as activity at end-systole and end-diastole (Figure 2a) (6,7,14), or from the peaks of a sinusoid (Figure 2b) (17,18,19) or any other suitable analytical curve (20) fit the data. The time-activity curves used for analysis typically have 40 or 50 millisecond time resolution and are usually constructed from list-mode data with 10 millisecond resolution. In this study, end-systolic activity is measured from time-activity curves during the 50 millisecond period immediately

*Single-probe studies (9,10,11) in which a single non-imaging collimated detector counts activity from the ventricle, are also used to generate time-activity curves.



Two method of computing left ventricular ejection fraction from time-activity curves:

a) Ejection fraction from ratios of end-systolic to end-diastolic activity

$$E.F. = \frac{\overline{ED}-\overline{ES}}{\overline{ED}},$$

where $\overline{ED}$ is the mean of the $\overline{ED}_K$, and $\overline{ES}$ is the mean of $\overline{ES}_K$.

b) Ejection fraction from amplitude of a sinusoid fitted to the data.

Figure 2

Two Methods of First-Pass Ejection Fraction Measurement

following subsidence of the T-wave (Figure 3) (21). Regardless of which method is used, subtraction of background activity (as observed in a properly chosen background area) is essential (17,22).

A somewhat more difficult problem is the estimation from radiocardiographic data of regurgitant fractions. Mitral regurgitant ejection fraction (M.R.E.F.) is defined as the fraction of end-diastolic left-ventricular volume ejected backward into the atrium through a leaky mitral valve. Aortic regurgitant filling fraction (A.R.F.F.) is defined as the fraction of end-diastolic volume flowing backward from the aorta into ventricle during diastole. Forward ejection fraction is the net fraction of end-diastolic volume ejected forward per beat.

The assumption of constant ventricular concentration during systole allows determination of E.F. regardless of the tracer input profile. However, the ease of measurement of mitral and aortic regurgitation is dependent on the form of tracer input. If tracer is injected as a bolus via a catheter directly into the atrium, or into the pulmonary artery, mitral regurgitation is manifested as a rise in total atrial tracer content during systole. If, on the other hand, the input to the left heart has a broad or dispersed distribution, as one would obtain from a peripheral I.V. injection, atrial tracer content may rise during systole because of mitral regurgitation or because of inflow of labeled venous blood, or both. This makes accurate visual interpretation of I.V. data for mitral regurgitation much more difficult, it not impossible.

There are several obvious reasons why quantitative noninvasive assessment of mitral and aortic regurgitation is of potential clinical importance. Mitral regurgitation is a common occurrence in several types of cardiac disease. It occurs in (a) rheumatic valvular disease, (b) left ventricular dilatation caused by congestive heart failure or primary myocardial disease, (c) ischemic dysfunction of the papillary muscle in coronary disease and in some myocardial infarctions (23). Conventionally, mitral regurgitation is diagnosed by indirect criteria such as heart sounds, and its extent is estimated angiographically during cardiac catheterization by observing the blackflow of dye from the ventricle into the atrium. Cardiac catheterization is an expensive and somewhat risky procedure for any patient, and often is considered too dangerous for patients with a recent myocardial infarction (M.I.) Since accurate measurements of mitral regurgitation have been difficult heretofore in recent M.I. patients, the frequency and clinical implications of temporary mitral regurgitation in acute M.I.'s is largely unknown.

Incompetence of the aortic valve is less than common than mitral incompetence but noninvasive assessment would be desirable in some cases. In aortic insufficiency, surgical treatment often carries a greater risk of morbidity and mortality than does medical management, up until the time in the course of the disease that left ventricular function begins to deteriorate. At that time, prompt surgery is called for to avoid progressive congestive heart failure (24). Periodic measurements of aortic regurgitation and ejection fraction by radiocardiographic means could help provide improved criteria for deciding when to intervene surgically.

Conventionally, aortic regurgitation is quantitated by comparing ventriculographically determined cardiac output with that derived from dye-dilution curves. This method is also used to measure mitral regurgitation. Hence, when both mitral and aortic regurgitation are present, only the total regurgitation can be measured by this means.

A reliable and noninvasive method for quantitating mitral and aortic regurgitation would clearly be useful for both clinical and research purposes. Though a successful method has been developed by Kirch, Metz, and Steele (14), it requires pulmonary wedge catheterization. Since several first-pass studies

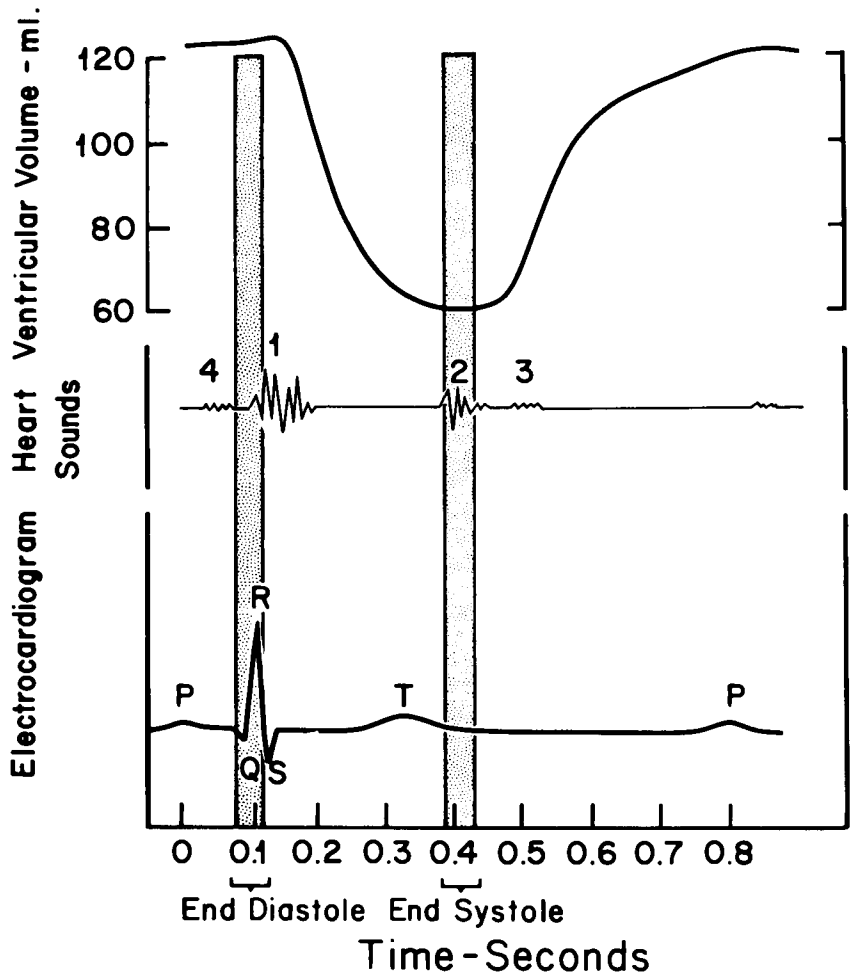


Figure 3
The relationship of electrocardiographic and phonocardiographic events with left ventricular volume, shaded bars represent intervals used for "gating" end-diastolic and end-systolic images.

are performed each day in our institution with I.V. tracer injection, we are developing a technique which may be used with I.V. injection.

THE MATHEMATICAL MODEL

A state variable approach has been used to formulate the estimation problem (Figure 4). The system under investigation, the cardiac pump, is represented as a linear time-invariant system operating on the state vector $\underline{X}(k)$, representing tracer concentration. Here k is a discrete time parameter and u(k) is a known scalar input. The state vector $\underline{X}(k)$ is observed by a process characterized by a linear observation equation. This observation includes random additive noise with zero mean, yielding the observation or measurement vector $\underline{Z}(k)$. In the problem considered here, the state variables are radioisotope tracer concentrations in the cardiac chambers, and the observation variables are points of the time-activity curves detected from areas of interest corresponding to the cardiac chambers.

The object of the procedure is estimation of the time-invariant parameter vector $\underline{p}$, and computing the observations to be expected from these trial systems. A conjugate gradient search algorithm is used to find the trial system which best matches the observations in some sense.

As an example of a more complex model with multiple phases, let us consider a general model of the central circulation. The possible pathways between the compartments of this general compartmental model are indicated in Figure 5 for the case including the more common shunts and the bronchial and coronary circulations. The model must account for delays and dispersive effects such as occur in the pulmonary circulation. In order to simulate the lungs in a central circulation model, one may use multiple serial mixing chambers with a delay, or other empirically-based transit-time distrubutions (11). In this system there are 14N individual flow parameters to be estimated, where N is the number of phases sampled during each cardiac cycle.

In the interests of simplifying the problem and reducing computing time, the number of flow parameters may be reduced by neglecting those pathways which are pathological (shunts) and those with neglibibly small flows. Significant shunts are relatively uncommon, and radiocardiographic methods exist for their detection and measurement (5,25,26). Routine screening of patient data using these methods could prevent inappropriate application of a shunt-free model to data from patients with shunts. The small farctions of cardiac output which flow through the coronary circulation (4 to 5 percent at rest [27]) and the bronchial circulation into the pulmonary veins (< 1 to 2 percent [31]) are neglible for our purposes. This reduces the number of interchamber pathways to six, and the number of parameters to 6N.

The number of flow parameters can be further reduced by dividing the cardiac cycle into only two phases, end-systole and end-diastole. This restriction has two adverse effects:

a) It eliminates the possibility of producing detailed ventricular volume curves, which can yield information on myocardial contractility and compliance (8), and

b) Only the small fraction of the time-activity curve corresponding to the periods of end-systole and end-diastole is usable, since other portions of the cardiac cycle are not modeled.

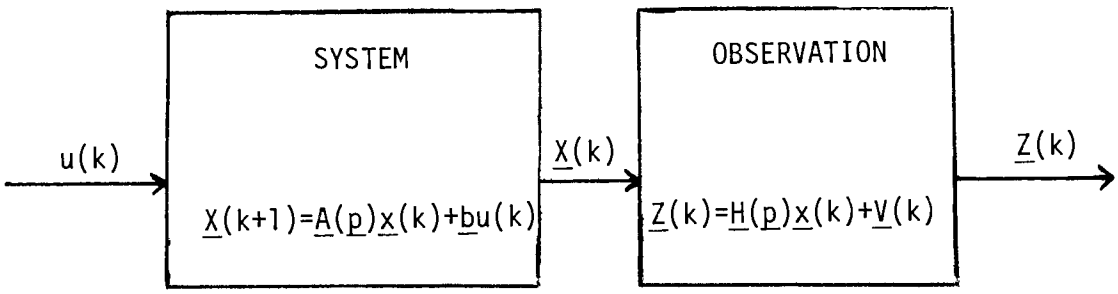


Fig. 4. Schematic of the linear system and linear observations of tracer activity. The parameters of interest are expressed as a vector $\underline{p}$.

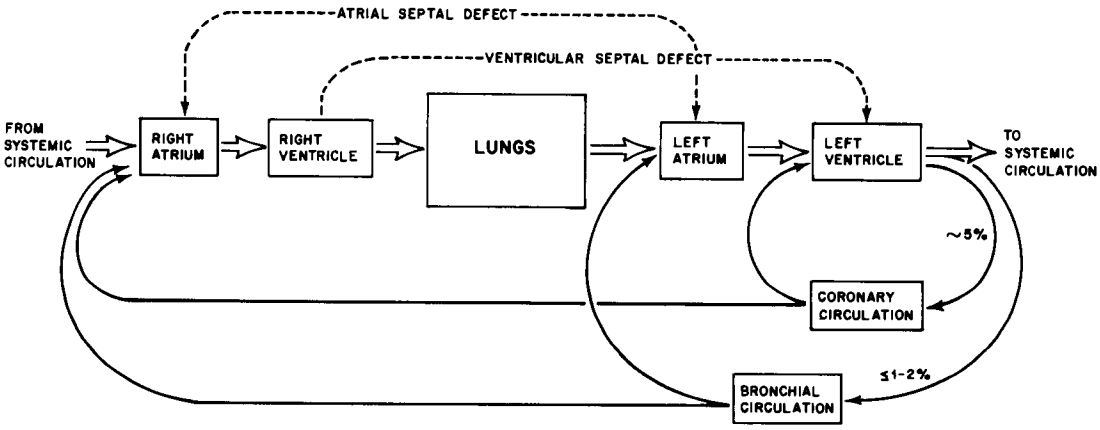


Fig. 5. Schematic of the flow of blood within the central circulation, including minor pathways (coronary and bronchial circulations) and shunts.

The model developed here does not attempt to represent the entire central circualtion. Only one side of the heart (atrium and ventricle) is modeled. As presently developed, it is applied to the left heart only, and the input function used is derived from delayed lung activity. (Similar two-phase models have been discussed by others (11,29), and Kirch and colleagues have applied a similar model in conjunction with z-transform methods to patient data obtained following impulse input via a pulmonary wedge catheter (14,16).

The model is expressed as a set of two different equations with atrial and ventricular tracer concentrations as the state variables.

Consider the schematized two-chamber system depicted in Figure 6. Blood containing tracer enters the system from the left and is pumped through the atrium and the ventricle. No valve is assumed at the inlet of the atrial chamber; the atrial-ventricular valve and ventricular-arterial valves may be incompetent. Note that the chambers are assumbed to pump alternately, and that the cycles are numbered such that ventricular end-diastole and the following end-systole are part of the same cycle. C(k) and V designate concentration and volume respectively, with the subscripts A and V designating atrium and ventricle respectively. For each compartment, the subscripts D and S represent the time of ventricular diastole and systole within the kth cardiac cycle. The term $\underline{u}(k)$ is the input concentration.

In the general case, where reverse flow from ventricle to atrium, and artery to ventricle is allowed, there are six intercompartmental flows, which we designate as follows:

f_{10} = reverse flow from atrium to venous input during diastole
f_{01} = forward flow from input to atrium during systole
f_{21} = reverse flow from ventricle to atrium during systole
f_{12} = forward flow from atrium to ventricle during diastole
f_{32} = reverse flow from output to ventricle during diastole
f_{23} = forward flow from ventricle to output during systole.

Note that these quantities are not really flows, but rather are volumes transferred into and out of atrium and ventricle.

Let F represent the net forward flow from any compartment to the next compartment during an entire cycle; that is, the forward stroke volume. Then we have

$$F = f_{01} - f_{10} = f_{12} - f_{21} = f_{23} - f_{32}. \tag{2}$$

This merely expresses the assumption that V_{AS}, V_{AD}, V_{VS}, and V_{VD} are constant with respect to the time index k.

Consider the quantity of tracer in the atrium at ventricular end-systole:

$$V_{AS}\, C_{AS}(k+1) = V_{AD}\, C_{AD}(k+1) = f_{21}\, C_{VD}(k+1) + [f_{01}-f_{10}]u(k) + f_{10}\, C_{AD}(k+1). \tag{3}$$

The last term represents atrial tracer that was ejected back into the venous compartment during the previous atrial contraction.

Since the tracer in each compartment is assumed to be well mixed at both end-systole and end-diastole, it follows that

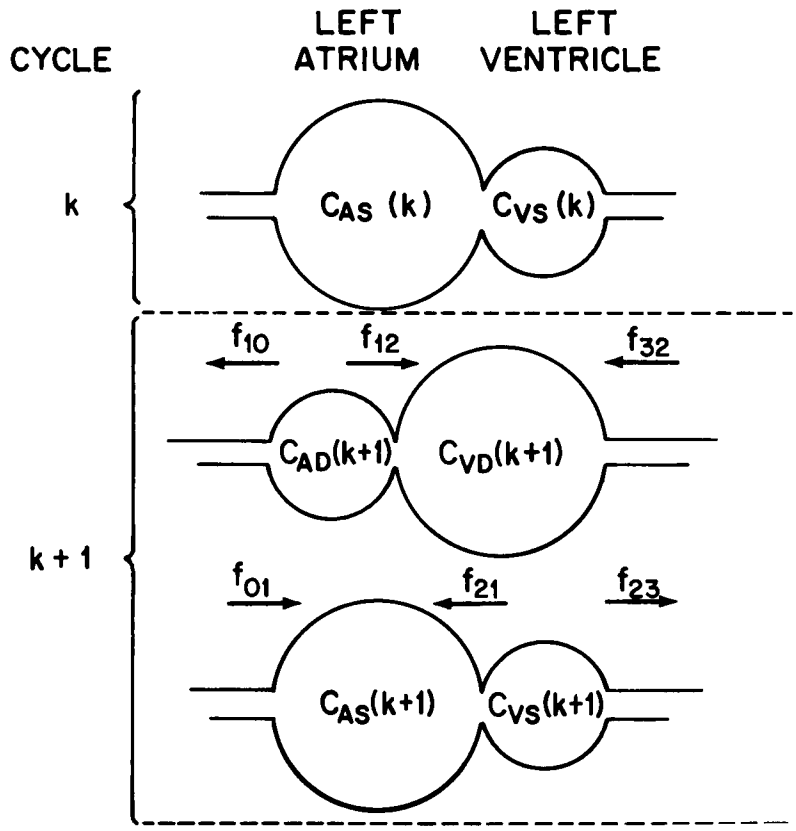


Figure 6

The simple two-stage model of left heart pump function used in parameter estimation algorithm. Symbols are explained in text.

$$C_{AD}(k+1) = C_{AS}(k) \quad , \tag{4}$$

$$C_{VS}(k+1) = C_{VD}(k+1) \quad . \tag{5}$$

Using equations (2) and (4), equation (3) can be written:

$$V_{AS}\, C_{AS}(k+1) = (V_{AD} + f_{10})\, C_{AS}(k) + f_{21}\, C_{VD}(k+1) + Fu(k). \tag{6}$$

The quantity of tracer in the ventricle during diastole is

$$V_{VD}\, C_{VD}(k+1) = V_{VS}\, C_{VS}(k) + f_{12}\, C_{AS}(k) + f_{32}\, C_{VS}(k) \quad ,$$

or

$$V_{VD} C_{VD}(k+1) = (V_{VS}\, f_{32})\, C_{VS}(k) + f_{12}\, C_{AS}(k). \tag{7}$$

Since we wish to use concentrations and not masses of tracer as state variables, we divide equations (6) and (7) by V_{VD}, and define the following parameters;

$$\eta_1 = \frac{F}{V_{VD}} = \text{Net forward ventricular ejection fraction} \quad ,$$

$$\eta_2 = \frac{f_{21}}{V_{VD}} = \text{Ventricular-atrial regurgitant fraction} \quad , \tag{8}$$

$$\eta_3 = \frac{V_{VD}}{V_{AS}} = \text{Ratio of maximum ventricular to maximum atrial volume.}$$

This allows the state equations to be expressed as

$$C_{AS}(k+1) = [1 + \eta_3(\eta_2-1)\ (\eta_1 + \eta_2)]\, C_{AS}(k) + \eta_2\eta_3\, [1-(\eta_1+\eta_2)]\, C_{VD}(k) + \eta_1\eta_3 u(k) \quad , \tag{9}$$

and

$$C_{VD}(k+1) = [1-(\eta_1+\eta_2)]\, C_{VD}(k) + (\eta_1+\eta_2)\, C_{AS}(k) \quad . \tag{10}$$

The three parameters appearing in the state equations (9) and (10) do not include one quantity of interest, the arterial-ventricular regurgitant fraction

$$\eta_5 \equiv \frac{f_{32}}{V_{VD}}.$$

Thus, the parameter estimation procedure cannot estimate η_5 on the basis of the state equations alone. If V_{VS}/V_{VD} is known, however, η_5 can be found indirectly from

$$\eta_5 = \frac{f_{32}}{V_{VD}} = 1 - \frac{V_{VS}}{V_{VD}} - \frac{F}{V_{VD}} - \frac{f_{21}}{V_{VD}} = 1 - \eta_4 - \eta_1 - \eta_2 \quad , \tag{11}$$

where η_4 is defined as

$$\eta_4 \triangleq \frac{V_{VS}}{V_{VD}} \quad .$$

In the case that the ventricle is observed entirely, η_4 may be estimated; that is, since we are assuming constant tracer concentration during systole, η_4 is just ventricular activity at end-systole divided by ventricular activity at end-diastole. Thus, if both end-diastolic and end-systolic are observed, η_4 appears in the set of observation equations (Equation (13) below), and may be estimated at the same time as η_1, η_2, and η_3.

IMPLEMENTATION OF THE METHOD AND THE ITERATIVE ESTIMATION ALGORITHM

The purpose of the model is to obtain estimates of the four parameters of cardiac function and chamber size defined in equations (8) and (11). Estimating the parameters is done by adjusting them until the model response, given the observed input, approximates the actual data in some sense. In this case the measure of closeness of fit is the weighted least-square cost function

$$J = \frac{1}{2} \sum_{i=1}^{3} \sum_{k=1}^{K} r_i(k)[z_i(k) - \hat{z}_i(k)]^2 \quad , \tag{12}$$

where $z_i(k)$ represents the observed data, and $\hat{z}_i(k)$ is the data as predicted from the model given trial parameter values. The subscript i indexes the three activity observations corresponding to C_{AS}, C_{VD}, and C_{VS} as defined previously; the index k, again, represents time in units of cardiac cycles.

Here $r_i(k)$ is a weighting factor representing an estimate of the relative statistical significance of the square deviation $[z_i(k) - \hat{z}_i(k)]^2$. In practice, the weights used are the inverses of the corresponding raw data points, which serve as estimates of Poisson variance.

Figure 7 summarizes the processing of the data. The unprocessed data is in the form of time-activity curves of high temporal resolution ($\leq$ 50 msec per data point). Figure 8 gives an example of left atrial and ventricular time-activity curves from a patient study. After the periods of end-diastole and end-systole are determined from a simultaneous electrocardiogram or phonocardiogram, background is subtracted. Determination of the weights $r_i(k)$, the delay δ between lung area of interest and left atrium, and the estimated observation coefficients A_1 and A_2 (Equation 13) complete the processing necessary for the iterative estimation algorithm to be applied.

One of the basic assumptions here is that left ventricular activity can be observed independently from left atrial activity, right ventricular activity and activity from major vessels. This condition can usually be met if the data is obtained from a modified left anterior oblique (30° left oblique, 15° caudal) view. In addition, at least a portion of the left atrium is isolated from other structures. The importance of this situation is that we can write the observation equations expressing the relationship of the observed activity to tracer concentration in a particularly simple form such as:

$$\begin{aligned} z_1(k) &= A_1X_1(k) + V_1(k) \quad , \\ z_2(k) &= A_2X_2(k) + V_2(k) \quad , \\ z_3(k) &= A_2\eta_4X_2(k) + V_3(k) \quad . \end{aligned} \tag{13}$$

Here the observed state variables $X_1(k)$ and $X_2(k)$ represent $C_{AS}(k)$ and $C_{VD}(k)$ respectively. The three observations here consist of atrial, end-systolic activity, and ventricular end-diastolic and end-systolic activities.

The constants A_1 and A_2 are observation scale factors representing the relative yield of counts per unit concentration for the atrial and ventricular areas of interest, where the corresponding factor for observation of the input is taken to be unity. This arbitrary choice is possible because only the ratios of concentration are of concern and determination of the absolute tracer concentration is not possible by the method described. Physically, these coefficients represent the differences in absorption, scattering, and detection geometry between the three areas of interest. The same coefficient, A_2, is assumed applicable to the ventricle during both end-systole and end-diastole.

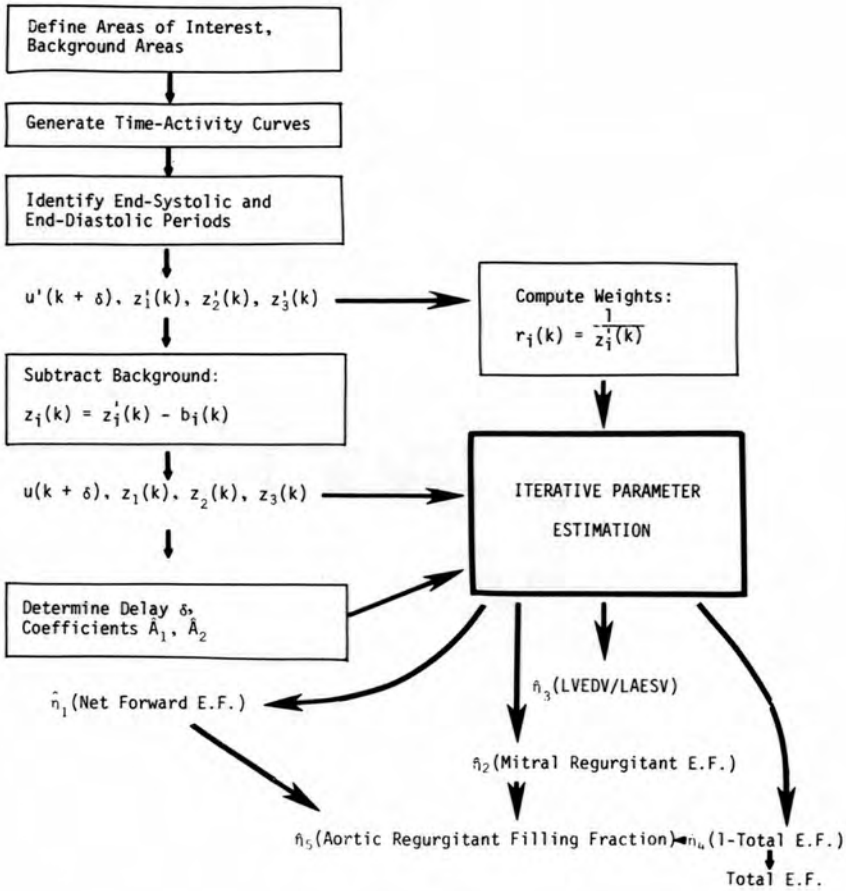


Figure 7. Flowchart of data processing for estimation of left ventricular ejection and valvular regurgitant fractions.

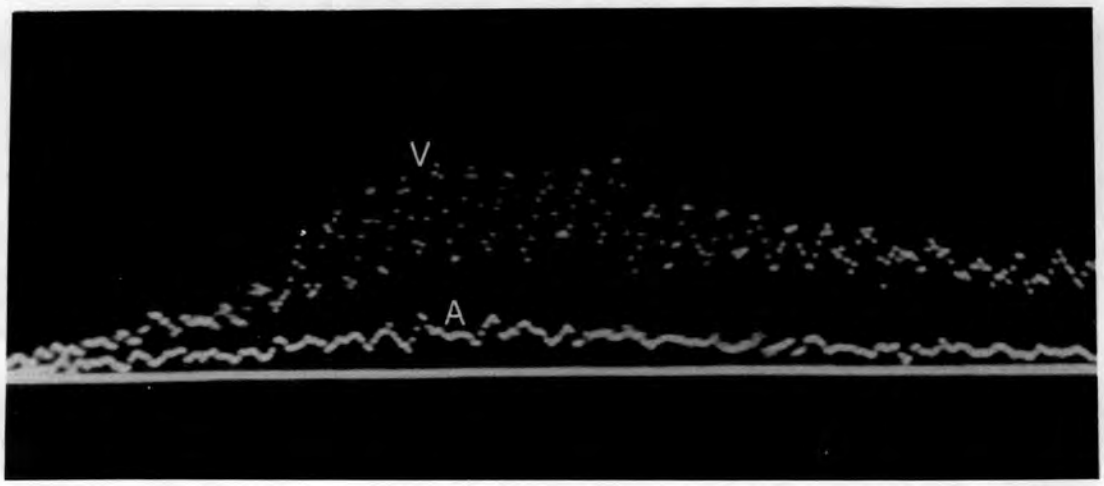


Figure 8. Time-activity curves from a patient study (MLAO view). Upper curve represents ventricular activity; lower represents atrial activity.

In practice, the scale factors are determined from observations of the areas of interest after the blood pool concentration has become essentially constant. At this time, comparison of atrial end-systolic activity (z_1) with input activity (u) gives A_1, comparison of ventricular end-diastolic activity (z_2) with input activity gives A_2, and an initial estimate of η_4 is given by the ratio of ventricular end-systolic and end-diastolic activities. (This estimate of η_4 may be used later as a starting value for the iterative estimation algorithm). If there were significant anatomical overlap of the chambers or "cross-talk" between them, the observation equations would contain additional unknown coefficients which would have to be estimated either by the algorithm or prior to its application.

For each observation, it is assumed that the appropriate background activity is determined and subtracted before using the estimation algorithm. Given that the chambers are adequately separated, which has been assumed, ventricular and atrial background subtraction requires only subtraction of lung activity as observed in an area adjacent to the chamber of interest.

In a study of the left heart, the input time-activity curve is measured from a region in the lung well separated from the cardiac chambers and large enough to be representative of the tracer concentration in the pulmonary blood pool. The input observations are smoothed (averaged) over a time period corresponding to a cardiac cycle. This reduces the noise in the input data, as well as elminating the effects of cyclic changes in pulmonary blood volume.

Note that the model input represents concentration in the pulmonary vein, just outside the left atrium. Since this region is difficult to observe, the input function is taken to be equivalent to the observed lung activity curve, but delayed by some time interval δ. The delay δ is estimated by observing the transit of some recognizable feature in the tracer concentration profile through the lung and then the atrium. In practice, this may be difficult.

SIMULATION STUDIES

The applicability of the model depends on conditions which may be conflicting. For instance, low frequency information is available only if there are significant spatial gradients in tracer concentration. However, the assumptions of instantaneous mixing in the chambers and "bolus flow" (flow that is uniform across the diameter of the blood vessels) become less accurate as spatial tracer gradients increase. That is, the behavior of the actual system more closely approximates that of the model, with its well-mixed chambers and bolus-flow vessels, if spatial tracer gradients are low; yet the lower these gradients become, the less sensitive the observations are to the transvalvular flows due to the lack of low-frequency information available. Clearly, it is important both for the model behavior to represent the system behavior, and for the transvalvular flows to be observable.

Apparently, mixing of tracer in the ventricle is in fact reasonably complete, since first-pass studies relying on this assumption give ejection fractions correlating well with cineangiographic ejection fractions (17). However, such questions as that of the sufficiency of low-frequency information in the data are perhaps best answered by simulation studies. Simulation can also help investigate the convergence properties of the minimization technique when applied to the data in question.

In order to test the performance of the method on cost functions closely resembling those generated from patient data, simulation stidues were performed using two sets of simulated curves, one set with a bolus input finction, the other with a broader input function of form $u(k) = Ck^2e^{-\alpha k}$ (where c and α are constants.

Table 1

SIMULATION RESULTS

Data Set	Actual and Estimated Values				
	η_1(TEF)	η_2(MREF)	$\eta_3(\frac{VVD}{VAS})$	η_4(1-TEF)	η_5(ARFF)
1*	*.55*	*.00*	*1.30*	*.45*	*.00*
1A†	.55	.00	1.30	.45	.00
1B††	.58	-.01	1.20	.45	-.01
2	*.30*	*.00*	*1.30*	*.70*	*.00*
2A	.29	.00	1.35	.69	.02
2B	.32	-.01	1.17	.70	.00
3	*.45*	*.10*	*1.10*	*.45*	*.00*
3A	.47	.10	1.05	.45	-.01
3B	.46	.11	1.08	.45	-.01
4	*.45*	*.00*	*1.30*	*.45*	*.10*
4A	.46	.00	1.25	.45	.10
4B	.47	-.01	1.2.	.45	.09
5	*.20*	*.00*	*3.00*	*.80*	*.00*
5A	.21	.00	2.74	.79	.00
5B	.21	.00	2.71	.80	.00
6	*.25*	*.30*	*.80*	*.45*	*.00*
6A	.24	.31	.86	.45	.00
6B	.30	.28	.64	.45	-.03

*Actual Value
†Estimates from bolus input
††Estimates from broad input

Simulation parameter values η_i were chosen to correspond to six normal and clinical situations. The parameters are intended to represent the following conditions: (1) normal total E.F. with no valvular regurgitation; (2) depressed E.F. with no regurgitation; (3) normal total E.F. with mild mitral regurgitation; (4) normal total E.F. with mild aortic regurgitation; (5) very low E.F. with enlarged left ventricle (as occurs in congestive heart failure); and (6) normal total E.F. with moderate mitral regurgitation and the left atrial enlargement which often occurs in mitral regurgitation. The 45-point simulated data curves were generated on a small computer (DEC PDP-8/I) directly from the model state equations (9) and (10) with bolus and dispersed inputs. A normally distributed noise term with Poisson variance and zero mean was added to the resulting curves to simulate the effects of Poisson noise, which occurs naturally in data taken from ramdom counts. The Poisson variance for a simulated data point was taken as the deterministic value of that data point.

The results of the simulation tests indicate that for parameter values within the ranges investigated, and for both input function profiles, that the parameters are identifiable in the sense that a global minimum of the cost function occurs very near the actual parameter values. Of course, due to the noisy nature of the data, the global minimum would not be expected to occur at the exact location of the true parameter vector. The course of a typical search is plotted in Figure 9.

Some convergence difficulties were encountered for cost functions which varied considerably more along some parameter directions than others. In these cases, the search tended to converge slowly and to stop before the blobal minimum was reached, even though the cost gradient was non-zero at the stopping point. It was always plain that the minimum had not been reached. Artificially "scaling" the gradient along the slowly-varying search direction usually allowed the conjugate-gradient search to converge to the global minimum. Unfortunately, the "scaling" required to accomplish convergence to the global minimum varied somewhat among the individual cost functions encountered. Thus multiplying the gradient vector by any single "scaling" vector acould probably not improve convergence in all cases. Similarly, it seems unlikely that actual scaling of the system model by linearly transforming the parameter vector (31) would improve convergence in all cases because this type of scaling is most efficient if the function before scaling is asymmetrical but quadratic in the parameters. Not only are the cost functions not quadratic in the parameters, but their shapes seem to vary considerably. Other minimization algorithms will be applied in the future in an attempt to improve convergence.

Note in Table 1 that estimates of η_2 and η_5 are sometimes small negative numbers. This corresponds to the non-physical situation of transvalvular flows counter to presumed pressure gradients, and is presumably due to the fact that the noisy cost function generated from the noisy simulated data has a global minimum displaced slightly from the "true" minimum.

PATIENT STUDIES

In a preliminary study of the method's applicability to patient data, the iterative estimation technique was applied to data acquired in radiocardiographic studies of four patients.

The studies were performed in the Coronary Intensive Care Unit of Parkland Memorial Hospital, Dallas, Texas, using a portable Anger camera (General Electric Portacamera) and dedicated minicomputer (G.E. Med II) for data acquisition and storage. The 45-point data curves were generated with the Med II as outlined previously; the search algorithm was performed using the DEC PDP-8/I.

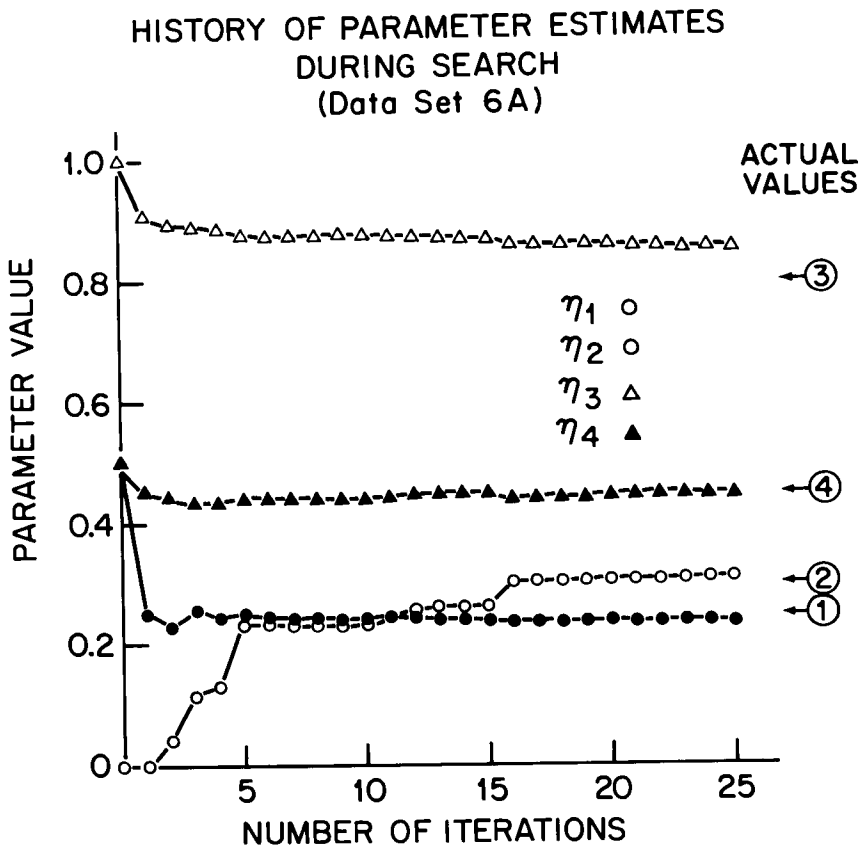


Figure 9. History of a typical parameter estimation search using simulated noisy data representing a system with known parameter values (indicated at right).

Two modifications were made to the model and estimation algorithm before it was applied to the patient data. A corrected model was used to account for physically occurring passive diastolic inflow, and appreciably large negative parameter estimates were prevented by imposing inequality constraints.

Passive Diastolic Inflow

The four-parameter model described previously neglected any passive diastolic inflow into the ventricle; that is, it assumed that the volume of blood flowing from atrium to ventricle during the cardiac cycle was transferred solely by atrial contraction during diastole. This is not true, since a variable fraction of the total flow from the atrium to the ventricle occurs passively as the ventricle relaxes prior to atrial contraction.

This passive diastolic inflow was introduced as an additional model parameter $n_o = \frac{f'_{12}}{V_{VD}}$, where f'_{12} is the volume of blood entering the ventricle after the mitral valve opens and before atrial contraction.

With introduction of passive diastolic inflow, the revised model state equations become

$$C_{AS}(k+1) = 1+\frac{V_{VD}}{V_{AS}} \quad \frac{F-f'_{12}+f_{21}}{V_{VD}} \quad \frac{f_{21}}{V_{VD}} - 1 + \frac{f_{21}}{V_{VD}} \cdot \frac{f'_{12}}{V_{VD}} \quad C_{AS}(k)$$

$$+ \frac{V_{VD}}{V_{AS}} \cdot \frac{f_{21}}{V_{VD}} \quad \frac{V_{VS}}{V_{VD}} + \frac{f_{32}}{V_{VD}} \quad C_{VD}(k) + \frac{F}{V_{VD}} \quad \frac{f'_{12}}{V_{VD}} \quad \frac{V_{VD}}{V_{AS}} \quad u(k) ,$$

$$= \{1 + \eta_3[(\eta_2-1)(\eta_1 + \eta_2) + \eta_0]\} C_{AS}(k)$$

$$+ \eta_2\eta_3[1-(\eta_1+\eta_2)]C_{VD}(k) + (\eta_1-\eta_2)\eta_3 u(k). \tag{14}$$

and

$$C_{VD}(k+1) = \frac{F+F_{21}}{V_{VD}} \quad C_{AS}(k) + \quad 1- \frac{(F+f_{21})}{V_{VD}} \quad C_{VD}(k) \quad =$$

$$(\eta_1 + \eta_2) \; C_{AS}(k) + [1-(\eta_1 + \eta_2)]C_{VD}(k). \tag{15}$$

Inequality Constraints

The second change was the imposition of inequality constraints in order to prevent parameter estimates from becoming significantly negative during the search. An additional term was added to the cost function J of equation 12; giving

$$J' = \frac{1}{2} \sum_{i=1}^{3} \sum_{k=1}^{45} r_i(k) \; [z_i(k) - z_i(k)]^2 + P \sum_{j=0}^{5} \eta_j^2 \; \mu[-\eta_j] ,$$

where P is an arbitrarily chosen large constant, and $\mu[\cdot]$ is the step function. By changing the cost gradient in regions of negative parameter values, an incentive is thus provided for the search algorithm to avoid these regions.

Patient Results

Two of the four patients studied (patients A and B) had recent acute myocardial infarction with no clinical evidence of mitral or aortic valvular

insufficiency. Patient C was studied with cardiac catheterization to assess the severity of his coronary artery disease (C.A.D.). No mitral or aortic regurgitation was observed. Patient D had clinical signs of mitral and severe aortic regurgitation, and both were demonstrated at cardiac catheterization.

Estimates of total and regurgitant fractions in the four patients are presented in Table 2, together with total E.F. measurements from gated blood pool (G.B.P.) studies and in patients C and D, total E.F. and total regurgitant fractions from contrast angiographic studies.

Although some mitral regurgitation may occur following myocardial infarction, the appreciable mitral regurgitant fraction estimated in patient B is probably excessively large. Nonetheless, the estimates of aortic regurgitant fraction compare favorably with clinical and angiographic evidence, and all estimates of total E.F. except that for patient C compare favorably with gated blood pool and angiographic results.

The estimates of total ejection fraction for patient C were 0.53 with the iterative method, and 0.77 by single-plane contrast angiography. In patient C, ejection fraction estimates from single-plane gated blood pool images were 0.57 in the modified left anterior oblique view, and 0.74 in the right anterior oblique view used also in computing patient D's contrast angiographic E.F.

Good agreement was found between I.P.E. and angiographic regurgitant fraction measurements in patient D, especially when regurgitant fraction was expressed as a fraction of stroke volume rather than as a fraction of end-diastolic volume (values in parentheses).

The patient results presented here must be considered preliminary for several reasons: (1) The scintillation camera used in these studies is less efficient and less suited to high count-rate first-pass studies than are newer cameras to be used in subsequent work. Higher count rates will allow more accurate measurement of volume by the "count-volume" method and more accurate delineation of ventricular boundaries. The camera presently used typically acquires 20,000 counts/second during the first pass of a 20 mCi tracer bolus. (2) The process of manually choosing the data points representing end-systole and end-diastole is subject to considerable error. An automated method for choosing end-diastolic and end-systolic data points on the basis of electrocardiographic or phonocardiographic signals recorded with the list mode data would produce more reliable results. (3) In the processing of patient data, it was found that 45 seconds of data is sometimes insufficient to establish a nearly equilibrated tracer concentration in the lungs, atrium, and ventricle. This near-equilibrium is required in order to accurately estimate the observation coefficients A_1 and A_2.

CONCLUSIONS

In this preliminary study with a limited number of patients, the iterative estimation technique successfully identified and estimated mitral and aortic valvular regurgitation in most cases. Also, the estimated left ventricular ejection fractions compared favorably with angiographic and gated blood pool estimates. Some caution is warranted, however. The number of patients in this study is small and a larger number of patients, some with mixed valvular insufficiency will have to be studied in order to completely assess the clinical utility of the proposed method. Nevertheless, this investigation represents initial attempts to accurately quantify forward and regurgitant fractions using a single noninvasive method coupled with a parameter estimation technique.

Table 2

ESTIMATES OF TOTAL AND REGURGITANT FRACTIONS IN PATIENTS

Preliminary results in four patients. Contrast angiographic studies of patients C and D were performed at cardiac catheterization. I.P.E. designates measurements by iterative parameter estimation. G.B.P. = gated blood pool. M.R.E.F=mitral regurgitant ejection fraction. A.R.F.F.=aortic regurgitant filling fraction.

Patient	Diagnosis	TOTAL E.F. I.P.E.	TOTAL E.F. G.B.E.	REGURGITANT FRACTIONS ANGIO	REGURGITANT FRACTIONS I.P.E. M.R.E.F	REGURGITANT FRACTIONS I.P.E. A.R.F.F.	REGURGITANT FRACTIONS I.P.E. TOTAL	Angiography Total
A 11/23/76	M.I.	.48	.38	NA[1]	.01	.00	.01	NA
B 11/23/76	M.I.	.46	.38	NA	.13	.00	.13	NA
C 2/14/77	C.A.D.	.53	.57 (MLAO) .74 (RAO)	.77(RAO) (2/15/77)	.06	.00	.06	0
D 4/1/77	AR, MR	.38	.37	.48 (3/30/77)	.00	.27 (.71	.27 .82)[3]	.39

[1]For patients A and B, angiographic results were not available.

[2]In patient C, single-plane gated blood pool (G.B.P.) E.F. was measured in RAO and MLAO views; contrast angiographic E.F. in RAO only.

[3]Regurgitant fractions in parentheses are fractions of stroke volume rather than fractions of end-diastolic volume.

REFERENCES

1. D. S. Berman, A. F. Salel, G. L. DeNardo, H. G. Borgren, and D. T. Morgan, "Clinical Assessment of Left Ventricular Regional Contraction Patterns and Ejection Fraction by High-Resolution Gated Scintigraphy," J. Nucl. Med. 16, 865 (1975).

2. H. W. Strauss, B. L. Zaret, P. J. Hurley, T. K. Natarajan, and B. Pitt, "A Scintigraphic Method for Measuring Left Ventricular Ejection Fraction in Man without Cardiac Catheterization," Am. J. Cardiol. 28, 575 (1971).

3. B. L. Zaret, H. W. Strauss, and P. J. Hurley, "A Non-Invasive Scintigraphic Method for Detecting Regional Ventricular Dysfunction in Man," N. Engl. J. Med. 284, 165 (1971).

4. C. B. Mullins, D. T. Mason, W. L. Ashburn, and J. Ross, Jr., "Determination of Ventricular Volume by Radioisotope Angiography," Am. J. Cardiol. 24, 72 (1969).

5. J. P. Kriss, "Acquired Cardiovascular Disease," p. 66 in Quantitative Nuclear Cardiology, Wiley, New York, 1975.

6. J. A. Parker, R. Secker-Walker, R. Hill, B. A. Siegel, and E. J. Potchen, "A New Technique for Calculation of L.V.E.F.," J. Nucl. Med. 649 (1972).

7. R. H. Secker-Walker, L. Resnick, H. Kunz, J. A. Parker, R. L. Hill, and E. J. Potchen, "Measurement of Left Ventricular Ejection Fraction," J. Nucl. Med. 14, 798 (1973).

8. M. V. Green, H. G. Ostrow, M. A. Douglas, R. W. Myers, R. N. Scott, J. J. Bailey, and G. S. Johnston, "High Temporal Resolution ECG-Gated Scintigraphic Angiocardiography," J. Nucl. Med. 95 (1975).

9. R. Folse and E. Braunwald, "Determination of Fraction of Left Ventricular Volume Ejected per Beat and of Ventricular End-Diastolic and Residual Volumes," Circulation 25, 674 (1962).

10. R. N. Pierson, Jr. and D. C. Van Dyke, "Analysis of Left Ventricular Function," p. 123 in Quantitative Nuclear Cardiology, Wiley, New York, 1975.

11. H. Parker, P. M. Weber, D. C. Van Dyke, H. Davies, P. Steele, and R. Sullivan, "Evaluation of Central Circulatory Dynamics with the Radionuclide Angiocardiogram," p. 67 in Cardiovascular Nuclear Medicine, H. W. Strauss and B. Pitt, Editors, C. V. Mosby, St. Louis, 1974.

12. S. Middleman, Transport Phenomena in the Cardiovascular System, Wiley-Interscience, New York, 1972.

13. H. Sandler, H. T. Dodge, R. E. Hay, et al,"Quantitation of Valvular Insufficiency in Man by Angiocardiography", Am. Heart J. 65, 501 (1963).

14. D. L. Kirch, C. E. Metz, and P. P. Steele, "Quantitation of Valvular Insufficiency by Computerized Radionuclide Angiocardiography", Am. J Cardiol. 34, 711 (1974).

15. D. Green, R. Carlisle, C. Grant, and I. L. Bunnell, "Estimation of Left Ventricular Volume by One-Plane Cineangiography", Circulation 35, 61 (1967).

16. P. Steele, D. Kirch, M. Matthews, and H. Davies, "Measurement of Left Heart Ejection Fraction and End-Diastolic Volume by a Computerized, Scintigraphic Technique Using a Wedged Pulmonary Arterial Catheter", Am. J. Cardiol. 34, 179 (1974).

17. H. R. Schelbert, J. W. Verba, A. D. Johnson, G. W. Brock, N. P. Alazraki, R. J. Rose, and W. L. Ashburn, "Nontraumatic Determination of Left Ventricular Ejection Fraction by Radionuclide Angiocardiography", Circulation 51, 902 (1975).

18. H. R. Schelbert, H. Henning, W. L. Ahsburn, J. W, Berba, J. S. Karliner, and R. A. O'Rouke, "Serial, Non-Invasive Measurements of the Left Ventricular Ejection Fraction Early and Late After Myocardial Infarction", Am. J. Cardiol. 38, 407 (1976).

19. D. B. Twieg, E. M. Stokely, R. W. Parkey, J. T. Willerson, and L. R. Nardizzi, "The Effects of Noise on the Determination of Ejection Fraction from Left Ventricular Time-Activity Curves", J. Nucl. Med. 18, 180 (1977).

20. F. C. Castellana, "Mathematical Modeling of the Central Circulation", p.202 in Quantitative Nuclear Cardiology, Wiley, New York, 1975.

21. R. M. Berne and M. N. Levy, Cardiovascular Physiology, Second Edition, C. V. Mosby, St. Louis, 1972, p. 76.

22. D. Van Dyke, H. O. Anger, R. W. Sullivan, W. R. Vetter, Y. Yano, and H. G. Parker, "Cardiac Evaluation from Radioisotope Dynamics", J. Nucl. Med. 585 (1972).

23. S. M. Ayres and J. J. Gregory, Cardiology: A Clinicophysiologic Approach, Appleton-Century Crofts, New York, 1971.

24. E. Rapaport, "Natural History of Aortic and Mitral Valve Disease", Am. J. Cardiol. 35, 221 (1975).

25. H. W. Strauss, "Detection and Quantification of Intracardiac Shunts", p. 128 in Cardiovascular Nuclear Medicine, H.W. Strauss and A.E. James, Editors, C.V. Mosby, St. Louis, 1974.

26. G. S. Freedman, "Radionuclide Angiocardiography in the Adult", p. 101 in Cardiovascular Nuclear Medicine, H. W. Strauss, B. Pitt, and A.E. James, Editors, C. V. Mosby, St. Louis, 1974.

27. A. C. Guyton, Medical Physiology, Fourth Edition, W. B. Saunders Co, Philadelphia, 1971.

28. A. P. Fishman, "Dynamics of the Pulmonary Circulation", Chapter 48 of Circulation, Vol. II, W. F. Hamilton and P. Dow, Editors, Handbook of Physiology, Section 2, American Physiological Society, Washington, D.C., 1973.

29. C. E. Metz, D. L, Kirch, and P. P. Steele, "A Mathematical Model for Determination of CArdiac Regurgitant and Ejection Fractions From Radioisotope Angiocardiograms", Physics in Medicine and Biology 20, 531 (1975).

30. W. J. McIntyre and J. P. Kriss, "The Normal Heart", p. 4 in Quantitative Nuclear Cardiology, Wiley, New York, 1975.

31. D. J. Wilde and C. S. Beightler, Foundations of Optimization, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967.

DISCUSSION:

The paper presented by Donald Twieg described a scintigraphic approach to quantitation of valvular regurgitation following a peripheral injection of a radionuclide bolus. By this concept it is imperative that an accurate estimate of the tracer input function be obtained from the lung region-of-interest. This input function is used in conjunction with background corrected left atrial and ventricular curves for iterative analysis by a conjugate gradient search algorithm which is applied to minimize a weighted least-squares cost function. This approach has been extensively tested on noisy simulated data with good results but only four patient studies were attempted to date.

In the discussion that followed this paper there was general acceptance of the conceptual and analytic approach embodied in this work. The importance of using pulsatile models rather than continuous flow models was emphasized. The main concerns of attendees focused on such practical problems as the ability to obtain an accurate input function from the lung region. Also, how do you know that the bolus is well mixed in the lung? Another concern was the ability to define regions-of-interest which separate the left atrium and ventricle as well as the right heart from the left heart. It was acknowledged that the lack of extensive evaluation in patients made it difficult to respond to these concerns at this time. Still, many people have successfully used count-volume analysis of cardiac data which carries the implicit assumption that the left atrium is well separated from the ventricle and that the right and left heart are separated as well.

SUMMARY OF PANEL DISCUSSION REGARDING THE EVALUATION OF LEFT VENTRICULAR FUNCTION: GATING VERSUS RCG TECHNIQUE

MODERATOR: DENNIS KIRCH
PANEL MEMBERS: MICHAEL GREEN, ANDREW TODD-POKROPEK, HORST GEFFERS AND DONALD TWIEG

The panel discussion began with several brief presentations regarding the technical and procedural aspects of gated scintigraphy using the Anger camera. Michael Green described the evolution from off-line, a posteriori gating using a large scale computer to a real-time, a priori system using a baseline minicomputer system. Using a hardware configuration consisting of a minicomputer central processor, keyboard terminal, display scope, and optional recording peripheral they are able to (1) develop a sequence of gated cardiac blood pool images in core, (2) use the R-R interval timing information to determine the beat-length distribution function, and (3) construct an averaged left ventricular volume curve over the average R-R interval. This approach uses a dual list-mode buffering technique so that the framing of the data can be synchronized within 10 milliseconds relative to the detection of an R-wave in the EKG signal. This system has had extensive application in performing rest versus exercise gated studies in which a single injection of 10 mCi of Tc-99m human serum albumin is used for both studies. While the development of the beat-length distribution function serves to indicate when a high percentage of irregular beats are occurring, it is not possible to eliminate these irregular beats by this real-time method without resorting to excessively long buffer lengths or experiencing extreme count losses. Still the availability of an immediate result is a definite advantage.

Andrew Todd-Pokropek discussed some aspects of a posteriori gating and how it might be useful to attempt to modify the data so that the same number of frames are developed between each pair of R-waves. This technique, called phase slicing, has been found to produce equivalent gated volume versus time curves in patients with regular hearts, but patients with irregular hearts cannot be adequately studied in this manner. It is generally conceded that variations in the R-R interval make it difficult to accurately predict the exact point of end-systole. That is to say, the left ventricular volume curve does not open and close like an accordian when the R-R interval is varied. Michael Green later pointed out that this statement is true for variations that occur as the rythmic heart rate is varied as well as for variations that occur on an arrythmic basis.

An interesting paper on the subject of sorting algorithms for EKG-gated cardiac scintigraphy was presented in another session by Mark Moore (see paper by Moore and Pizer in this publication). This paper presents a sorting technique in which list-mode data is presorted into equivalence classes prior to transferal to disk storage. Beat number information is merged with the list mode data as well. Several advantages are realized by this method. It can reasonably be performed on dedicated minicomputer systems of the type widely available in Nuclear Medicine today. By keeping track of irregular beats the list mode data corresponding to these beats can be eliminated during the reformating of the data to frame mode. The data can be sorted with a minimum of disk head motion within a few minutes following completion of the study.

This approach results in excellent performance characteristics which could potentially allow near real-time gating to be performed without sacrificing some of the advantages of off-line gating.

Dennis Kirch discussed a tomographic technique using a large-field camera (see paper by Vogel, Kirch, LeFree, et.al. in this publication) that could potentially be applied to first transit cardiac flow studies as well as gated studies. This would eliminate the problem of limited viewing angles for gated scintigraphy. He then described the application of a simple precordial scintillation probe approach to the accurate measurement of left ventricular volume as function of time (see paper by Frischknecht, Kirch, Quinlan, et al.

In the discussion that followed the presentations by the panel members it was clear that the group in attendance had little appreciation or interest in the fact that a single scintillation probe can produce left ventricular volume data which is extremely accurate when compared with frame-by-frame analysis of contrast cine ventriculograms. The discussion was almost entirely devoted to technical details related to gated scintigraphy and first transit camera studies. Concern was expressed with regard to the ultimate patient radiation dosage required by the different methodologies for performing rest and exercise studies. Multiple first transit cameras studies were conceded to result in the highest dosage (about 20 mCi of a suitable Tc-99m compound is required for each study). The gated scintigraphic technique allows both the rest and exercise studies to be performed with a single injection of 10 mCi of Tc-99m HSA. The single probe technique requires only 2 mCi per study but cannot provide regional wall motion information.

Some basic questions regarding the validity of left ventricular volume versus time data produced by any of the radionuclide methods were raised. Very little effort has been devoted to comparison of the time averaged curves produced by gated scintigraphy with frame-by-frame analysis of the contrast cine ventriculograms on the same patients. Michael Green has made measurements in animals with cardiac flow metering devices and compared this information with the more significant temporal events such as onset and end of systole as determined by gated scintigraphy with good agreement. Dennis Kirch related that extensive correlations between cine and single probe volume data show excellent agreement and that aortic arch pressure has been recorded simultaneously with the volume data to establish that the onset and end of systole are accurately reflected in the single probe volume curves. Still there are many serious questions to be answered before volume curves produced by radionuclide techniques can be accepted as valid. Effects of time varying background, sensitivity variations in the detected count rate as the distance from the heart to the detector varies and contributions from adjacent cardiac chambers were all expressed as factors which could potentially distort scintigraphically determined volume data.

It was stated that the casual observer is definitely confused by the wide variety of radionuclide techniques that are being proposed for making similar or identical measurements of cardiac performance. Some of the attendees felt that such a degree of disparity and lack of unity will make future progress more difficult and that some standards should be adopted and retained for long term studies. The panel recognized that such diversity exists but reminded the attendees that clinical versus investigational applications and differences in the operating characteristics of the available equipment are factors which play a strong role in determining the methods of choice. Adoption of standardized techniques for the performance of long term studies must be done individually or by members of those groups which choose to perform collaborative studies. It was further emphasized that even the most widely applied radionuclide cardiac studies such as ejection fraction measurements should be

validated independently by each laboratory or even each individual prior to application for long term studies. Some universal truths regarding radionuclide measurements of cardiac performance are now recognized but it will be a few years before a "text book" form of nuclear cardiology emerges.

QUANTITATIVE STUDIES WITH THE GAMMA-CAMERA : CORRECTION FOR SPATIAL AND ENERGY DISTORTION

F. Soussaline, A.E. Todd-Pokropek, C. Raynaud
Commissariat a l'Energie Atomique - Departement de Biologie
Service Hospitalier Frederic Joliot - 91 406 Orsay - France
Department of Medical Physics - University College Hospital, London-England

ABSTRACT

The gamma camera sensitivity distribution is an important source of error in quantitative studies. In addition, spatial distortion produces apparent variations in count density which degrades quantitative studies. The flood field image takes into account both effects and is influenced by the pile-up of the tail distribution.

It is essential to measure separately each of these parameters. These were investigated using a point source displaced by a special scanning table with two X,Y stepping motors of 10 micron precision. The spatial distribution of the sensitivity, spatial distortion and photopeak in the field of view were measured and compared for different setting-up of the camera and PM gains.

For well-tuned cameras, the sensitivity is fairly constant, while the variations appearing in the flood field image are primarily due to spatial distortion, the former more dependent than the latter on the energy window setting. This indicates why conventional flood field uniformity correction must not be applied.

A correction technique to improve the results in quantitative studies has been tested using a continuously matched energy window at every point within the field. A method for correcting spatial distortion is also proposed, where, after an adequately sampled measurement of this error, a transformation can be applied to calculate the true position of events. The knowledge of the magnitude of these parameters is essential in the routine use and design of detector systems.

INTRODUCTION

Non-uniformity of a gamma camera is an important source of error both in quantitative and qualitative studies. Many workers have considered that the principal source of error is due to variations in sensitivity, and have measured this effect by looking at the flood field image [1-3].

Previous studies have clearly shown that using the flood field image to correct for non-uniformity by "matrix division" increases the error in quantitative studies [4,5]. The reason for this error is that the flood field does not measure variations in sensitivity alone, but represents a combination of variations in sensitivity, spatial distortion and tail effects.

The aim of this paper is firstly, to measure separately the various components of non-uniformity and, secondly, to establish methods for their elimination.

THE CAUSES OF NON-UNIFORMITY

The non-uniformity in response of a conventional gamma camera has, from a theoretical point of view, three causes :

a) Variations in the sensitivity distribution, i.e. the variation in count rate as a function of position,

b) the spatial distortion, or non-linearities,

c) the tail effects, i.e. variation in the shape of the point spread function [6].

This third cause of error has been measured and is negligible, with pile up going from 0.1-1.1% with scattering medium thickness of 0-15 cms, respectively.

POINT SOURCES AND EXTENDED SOURCES

It is usually assumed that a flood field image gives results equivalent to measuring the count rate for a point source placed at every discrete position in the field of the camera. This is completely false and quite different results are usually obtained in the two cases.

An extended source is indeed identical to a collection of point sources, but it must not be imagined that the data at a point in images from an extended source correspond to values that would be obtained for a point source at the corresponding position. It is possible to reconstruct the results obtained for an extended source from a set of point source images, but not, in our opinion, the inverse since the three effects of sensitivity, spatial distortion and tails seem to be inextricably intertwined. Thus, it is essential to measure separately these three components. Scanning a point source in the field of view of the camera is a very direct and straight forward technique for performing these measurements.

METHOD

In order to scan a point source a special scanning table was constructed, as illustrated in Fig. 1. It comprised two perpendicular racked 'X' bars of optical quality. By means of two stepping motors a lead pot may be placed anywhere in a field of 400 x 400 mm in steps of 0.1mm, with a precision of better than 20 microns. The stepping motors are controlled directly by the computer data processing system, attached to the camera.

Two cameras, a TOSHIBA (delay line) GCA 202 and a NUCLEAR CHICAGO Pho gamma III both connected to an INFORMATEK SIMIS III data processing system, were tested with this configuration. Fig. 2 shows a block diagram of a portion of the interface. For the TOSHIBA gamma camera the window selector has been modified and the 'Z out' signal encoded. Three isotopes were used being 197Hg(70 keV), 99mTc(142 keV) and 131 I(360 keV).

For each position (I,J), the count rate, the energy spectrum and the center of gravity in camera coordinates were determined. Decay correction was performed. This operation was repeated at every point on a regular lattice of points in table coordinates. The count rate for each measurement was about 2000 cps.

The precision of the determination of the center of gravity was studied. Limited optimally by the integral linearity of the gamma camera-computer ADC's (of the order of 0.1%) reproducibility of the order of 0.1% was obtained. This corresponds to a precision of about 1/5 of a pixel when the matrix is displayed in 256 x 256, or of the order of 0.3mm for the large field of view TOSHIBA. This precision was obtained from a sample of 5000 counts, but did not vary rapidly as a function of total counts. For 2 K to 100 K counts very little difference was observed. Typically, all the measures described in this paper were made with 5 K counts.

The size of the point source did not significantly affect the precision of this measure and although a highly collimated point source was used originally, this was later replaced by an 'uncollimated' point source of ∿1.5 cm in diameter. All measurements were made with the collimator in place, and at 5 cm separation between pot and collimator.

The count rate was determined by collecting a fixed number of counts and determining the time with a precision of 0.01 sec. The field of the camera was delimited as that area in which the count rate did not fall below 250 cps.

In order to relate table coordinates (I,J) and camera coordinates (X,Y), a calibration was performed. Measurements of the center of gravity of the point source at four different points with 100 K counts were made and the five parameters : angle, scale X, scale Y, origin X, origin Y were determined. Using these values, coordinates in (I,J) could be transformed into (X,Y) and vice-versa, with a precision of the order of 0.5%. This transformation was rarely used and the experimental results presented here are independent of this calibration.

Various sampling sizes were used being from one point per mm up to one point per 20 mm. The total time varies strongly as a function of sampling size, being from ∿20 minutes for a very coarse grid to several hours for a very fine grid (when the isotope is decaying).

Rather than storing the complete spectrum at every point, the peak value was determined with a precision of 1/256 and stored.

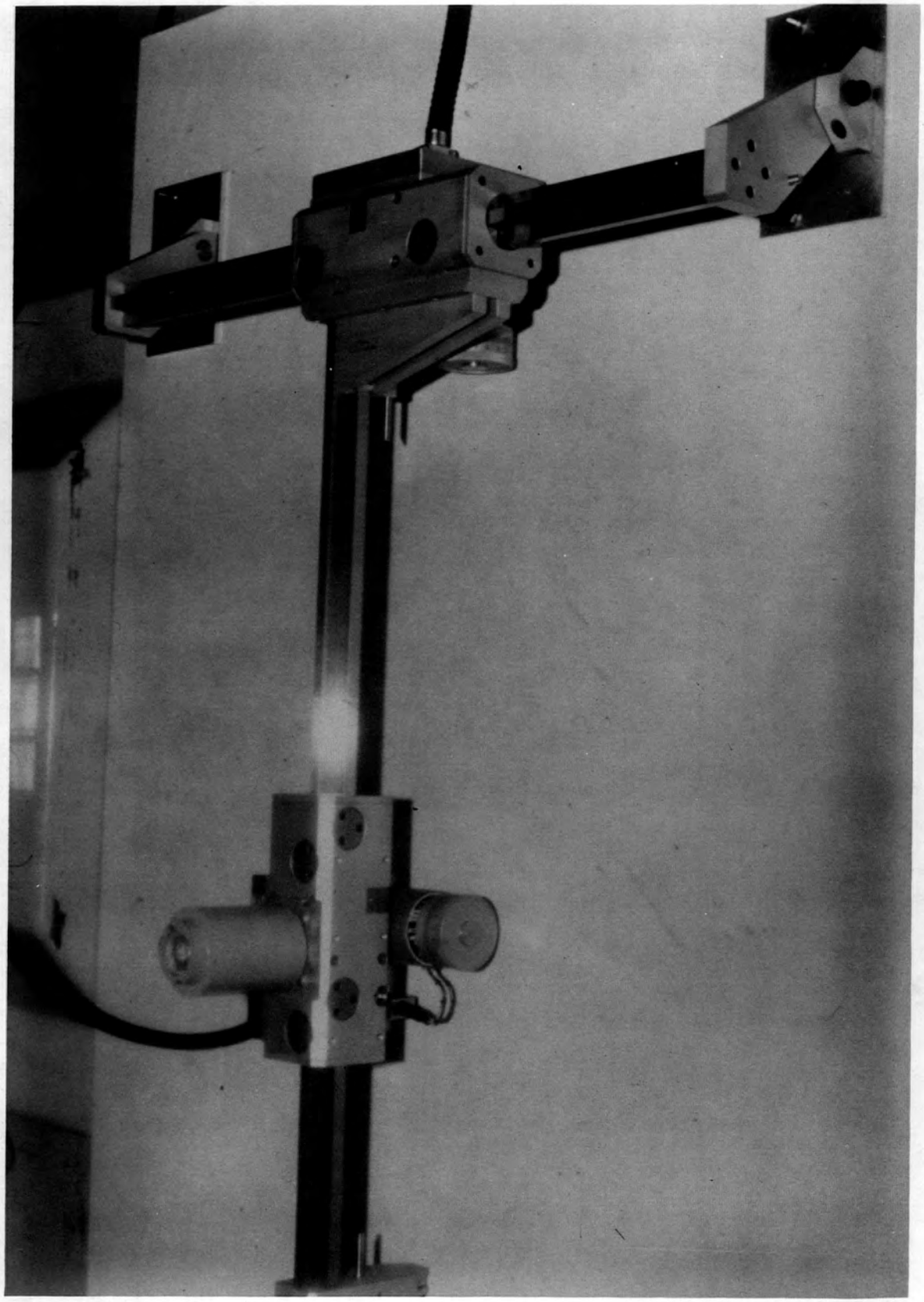


Figure 1: X ,Y scanning table with two stepping motors.

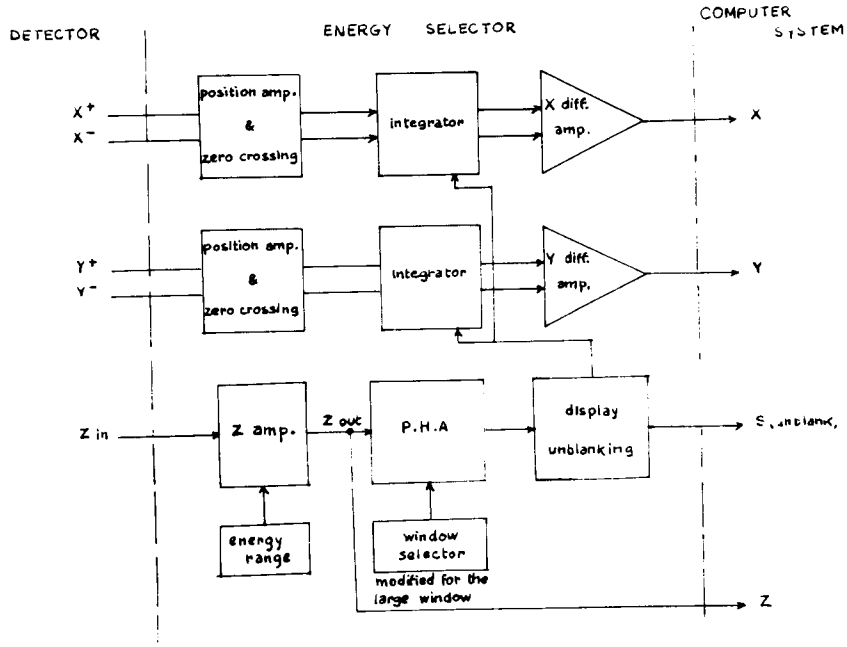


Figure 2 : Block diagram of a portion of the interface.

The entire operation was fully automated and no operator intervention was required.

RESULTS

Spatial Distortion Magnitude

Although it was originally felt that spatial distortion could be measured by determining the distance between the true position of a point source and the observed position, it is considered that this parameter is much less important than that of scale. By 'scale' the concept of the ratio between a certain real distance, and that same distance as measured is implied. It is a change in 'scale' which alters the 'apparent' count density at a point. Fortunately, 'scale' is also much easier to measure. Although very large values for spatial mispositioning of the order of one centimeter are observed at the edges of the field of view, in the center, spatial displacement is of the order of a millimeter. It is difficult to be very precise about the amplitude of this error since it is very dependent on the positional calibration (which is dependent on the selection of the calibration points). For example, there can be no 'spatial displacement' at the calibration points themselves by definition since the true and observed coordinates have been forced to coincide. Thus these measurements are rather arbitrary, unlike the determination of scale which is precise. This problem is further discussed in the 'Correction of Spatial Distortion' section.

If it is recalled that the data as stored consist of a set of values for (X,Y) observed position (I,J) true position, T time to collect N counts, Z the peak in energy etc....., a simple method of determining 'scale' may be used. Adjacent points (in both the horizontal and vertical direction) are equally spaced. Thus, the amplitude of the vectorial distance between adjacent pairs of (X,Y) values corresponds to the measurement of the observed distance for a constant real distance, the ratio of the two being 'scale'. 'Scale' can be determined regularly sampled in (I,J), and irregularly sampled in (X,Y) gamma camera coordinates.

However, if the assumption is made that this parameter is slowly varying with respect to the element size of the regular grid on which it is to be displayed, regularisation in (X,Y) may be achieved simply by coarse interpolation. When several observations exist for one element, the average value was used. When no sample value exists, the value is derived by interpolation from existing horizontal and vertical neighbours.

Although spatial distortion is truly a vector, 'scale' is a scaler. It corresponds to the real bin size of each element in the displayed matrix. It is therefore convenient to use as a display matrix the same 64 x 64 grid used for gamma camera images. For a constant count density, variations in bin size will create apparent variations in count density. It is this effect which prevents the flood field image from being a true representation of the sensitivity distribution.

Fig. 3 shows a 64 x 64 representation of 'scale' where brightness is directly proportional to observed distance divided by real distance, for a well adjusted camera. In order to estimate the effect of this in terms of flood field image, the reciprocal must be used. The greater the observed distance, the smaller the bin size and the lower the apparent count density. Hereafterwards, 'scale' will be displayed as true distance

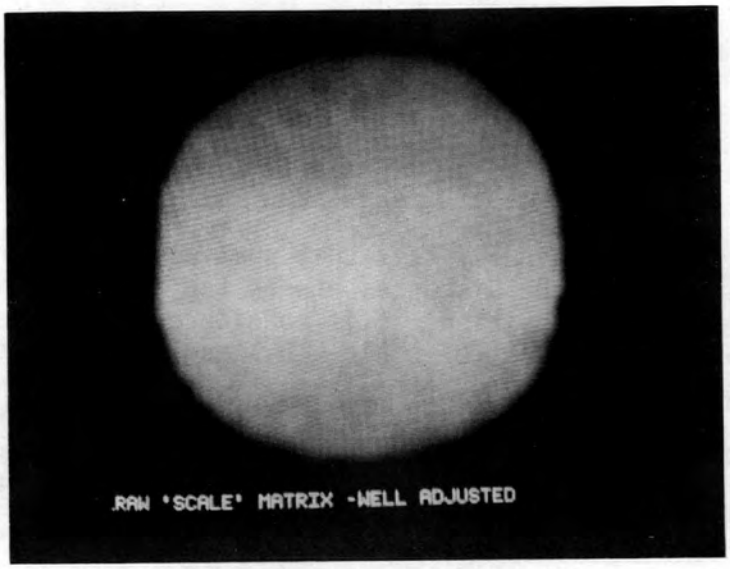


Figure 3 : Raw 'scale' matrix for a 'well' adjusted camera.

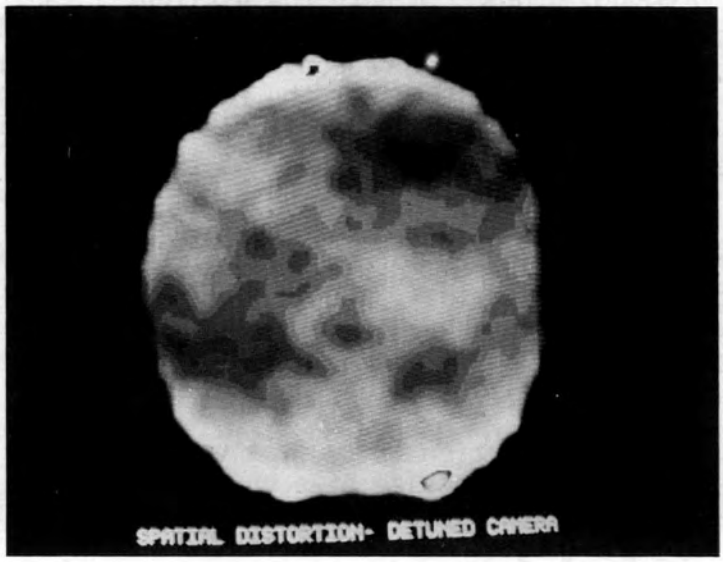


Figure 4 : Spatial distortion matrix for a detuned camera.

divided by observed distance where, as shown in Fig. 4 for a badly tuned camera. Holes in this matrix correspond to regions of decreased bin size which could give rise to lower values of count density. These matrices can be compared to the flood field as shown in Fig. 5.

The numeric values observed of the apparent count rate changes due to spatial distortion are very revealing. On the edges, values of up to 60% above the minimum were found. In the center, within the usable field, variations of ± 10% were observed. This seems to contradict the conclusions of V. Brookmann, who stated that spatial distortion was not significant [7].

It is interesting to look at the different images of spatial distortion obtained with various sampling sizes for the point source displacement. Fig. 6 shows three such images corresponding to a 3.5 mm, 10.mm and 20.mm sampling. Although the interpolation has created diffuse features, the profiles across the 'hole' show that the contrast is approximately the same in all three images.

The variation in spatial distortion (and sensitivity) as a function of time was determined. Both short term and long term changes (over a period of days) with the TOSHIBA camera were smaller than the calculated and observed measurement error. This seems to indicate that the use of this scanning technique, where such measurements can take substantial time, is valid.

Sensitivity

Using the table, a regular lattice of values for the time to collect a fixed number of counts was obtained. These data were converted into a regular 64 x 64 matrix in camera coordinates by the same process of interpolation as described above. Again the reciprocal of this matrix was used such that each image values represented count rate at that point.

Fig. 7 shows the sensitivity image for a well tuned camera and may be compared to Fig. 8 which shows the effect of detuning a photomultiplier.

As has been previously stated [4,5], the sensitivity response of a well tuned camera is remarkably flat. Fig. 9 shows a profile through such a sensitivity matrix showing variations of less than 3.5% of the mean value. Note there is no increase in sensitivity at the edges.

Spatial Distortion, Sensitivity and the Flood Field

It is interesting and revealing to compare the observed distributions of spatial distortion, sensitivity, and the flood field. Fig. 10 shows one such comparison for a camera with a detuned PM and a well centered 15% energy window. It may be observed that the flood field is more like the spatial distortion matrix than the sensitivity matrix. This confirms the hypothesis that since a camera is basically uniform in sensitivity when well adjusted, the flood field shows primarily variations in the bin size due to spatial distortion. Clearly, conventional matrix multiplication uniformity correction is undesirable and will tend to increase the error in quantitative measurements. The edge packing effect seen with the flood image is due entirely to spatial distortion. An image obtained with a Searle H.P. camera, showing the distortion matrix with hot spots due to the PMs,is illustrated in Fig. 11.

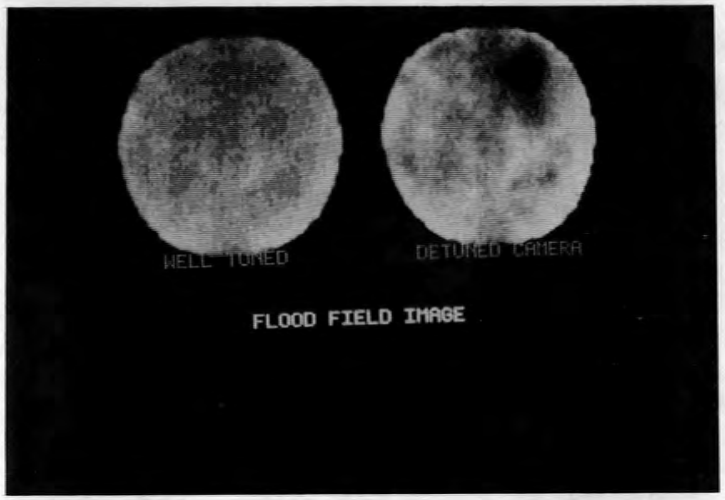
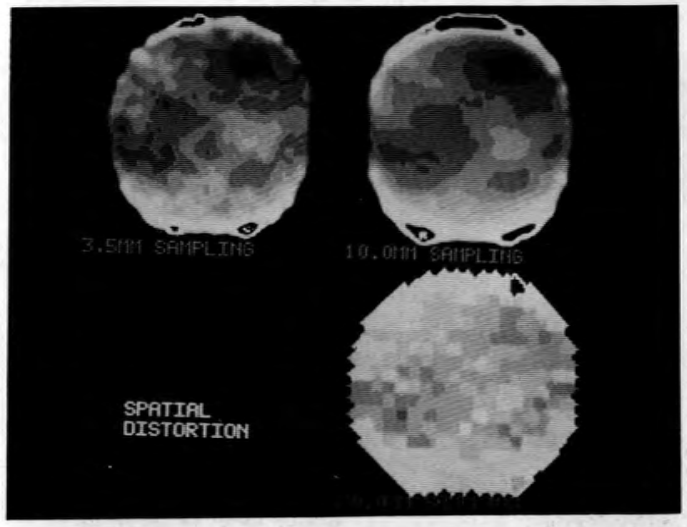


Figure 5 : Flood field image obtained under the same conditions for a well tuned (left) and a detuned (right) camera.



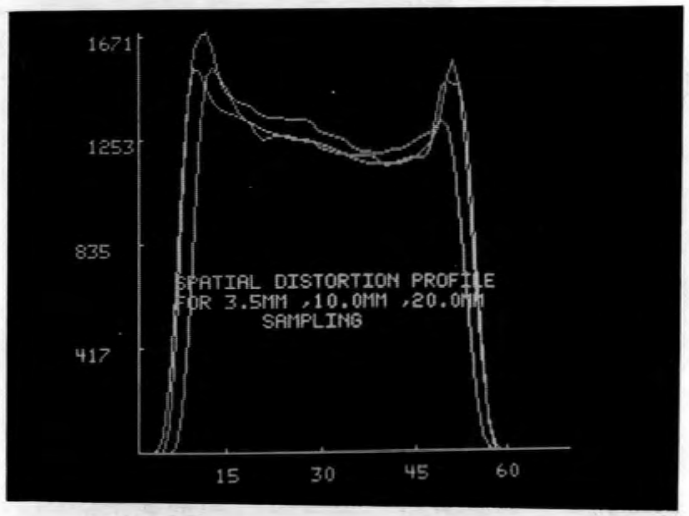


Figure 6 : Spatial distortion matrixes for different sampling sizes ant the corresponding profiles.
Upper left = 3.5 mm, Upper right = 10.0 mm,
below = 20.0 mm sampling.

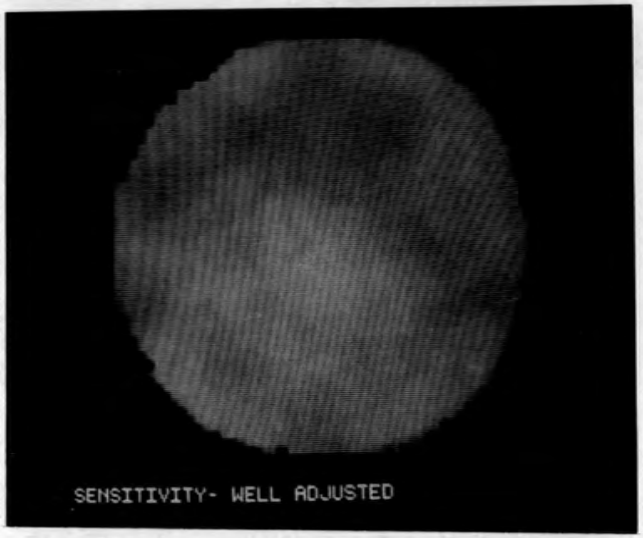


Figure 7 : Sensitivity matrix for a well tuned camera.

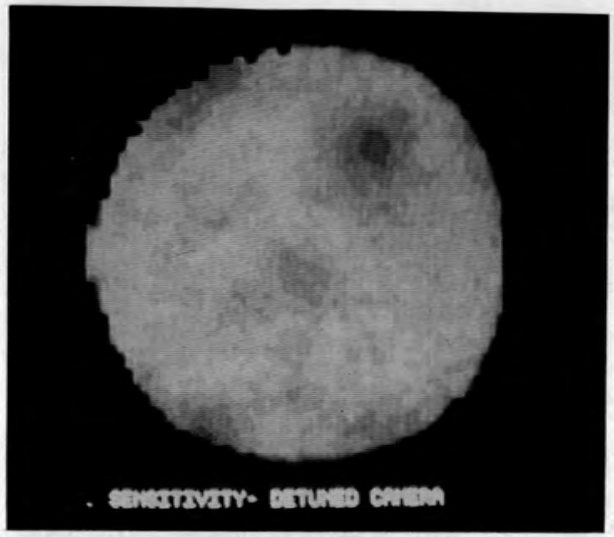


Figure 8 : Sensitivity matrix for a detuned camera.

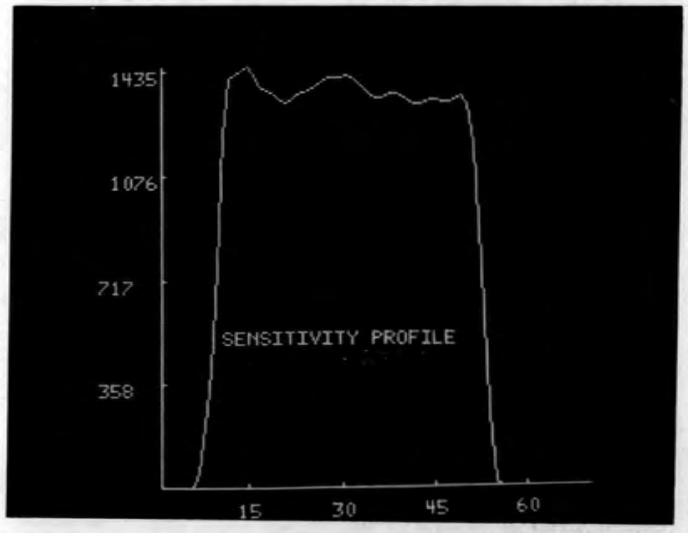


Figure 9 : Profile corresponding to figure 7.

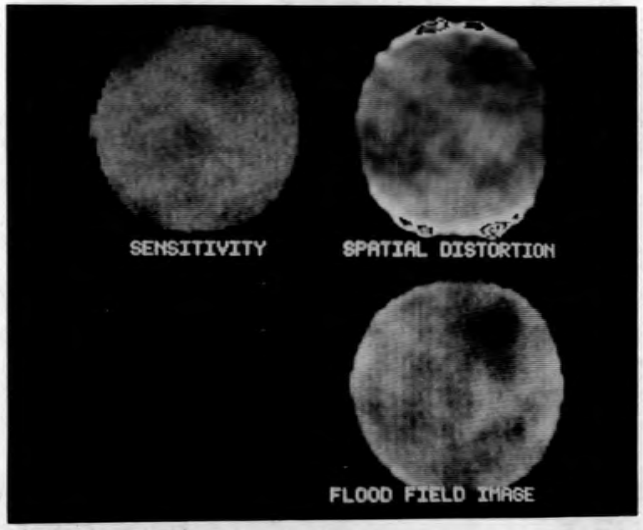


Figure 10 : Spatial distortion (upper right), sensitivity (upper left) and flood field (below) images for a detuned camera and a well centered 15% energy window.

Although, the spatial distortion matrix and flood field are much more similar, this does not imply that there is no sensitivity component in the flood field. As shown in Fig. 12, the flood field can sometimes have a very significant feature created by the sensitivity distribution, the hot spot (at 2 o'clock) due to a detuned PM is present in the sensitivity and flood field images, but not clearly visible in the spatial distortion image. Here the gain of the PM was increased and the window was set high.

Sensitivity as a Function of the Energy Window

With a narrow 15% energy window, it is well known that different results are obtained for flood field image depending on whether the window is high, low or well centered with respect to the photopeak. The same effect has also been observed for the sensitivity matrix. A single PM was detuned such that the photopeak was 4% lower than for the other PM's. Fig. 13 hows the sensitivity matrix for a well centered window, a low energy window centered in the detuned PM photopeak, and a window raised by the same amount. For the well centered and high windows there is a hole corresponding to the detuned PM and as might be expected from physical considerations, there is a hot spot for the low window image. The sensitivity distribution is clearly very dependent on the setting up of the window.

Fig. 14 shows an even more surprising result comparing the flood field, the sensitivity distribution and the spatial distortion for this last low window setting of the PHA. The flood field and spatial distortion both show cold spots, whereas the sensitivity distribution shows a hot spot. Thus the flood field seems to indicate a cold spot at that point where actually the camera is more sensitive. This shows strikingly why flood field uniformity correction must not be applied.

Energy Distortion

Ideally, it is considered that from physical considerations for a constant sensitivity over the field of the camera the gain of all the PMs must be tuned such that the energy spectra are all identical. This may be performed by placing a point source in front of each PM and adjusting the gain using the same technique described above. Having set up the camera in this manner, then the peak in energy for a camera was determined as in the 'Spatial Distortion Magnitude' section, and displayed as a 64 x 64 matrix. Fig. 15 shows the energy distribution compared to the corresponding sensitivity distribution as shown in Fig. 8. It may be noted that there are considerable similarities. This indicates that it might be useful to correct the variation in energy as a function of position.

Sliding Energy Window Correction

Previous work has been performed using a sliding energy window such that the center of the window is matched to the peak in spectrum at every position. This determined the energy matrix by sampling 8x8 [8] and 4x4 areas [5]. In this study, the Z energy matrix was determined at every point in a 64 x 64 matrix. Using the value obtained from the first study to center correctly the 'digital' window, the sensitivity and spatial distortion matrices were again determined in a second study. It must be noted that this 'digital' window, which was chosen to be 10keV/wide had a constant width at each position. To perform an accurate correction, the relative width should be

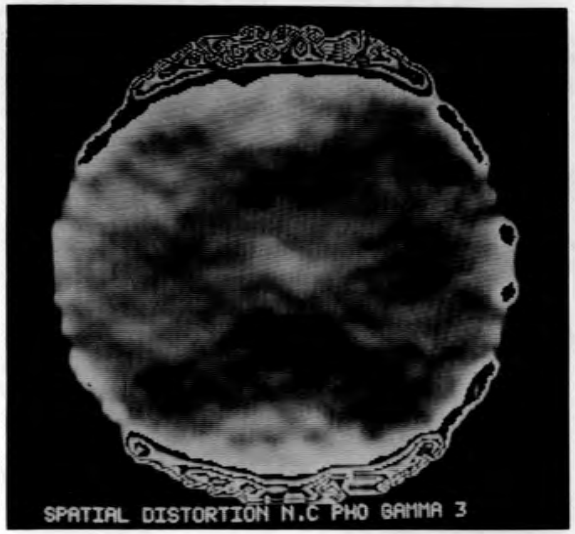


Figure 11 : Spatial distortion matrix for a Searle H.P. camera.

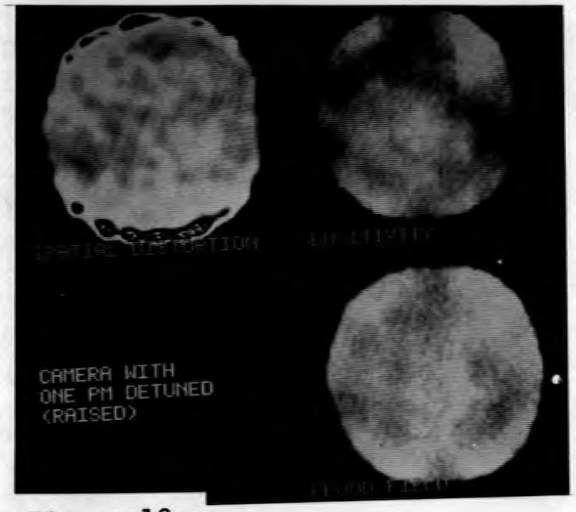


Figure 12 :
Spatial distortion (upper left), sensitivity (upper right) and flood field (below) images for a camera with a detuned (raised) PM.

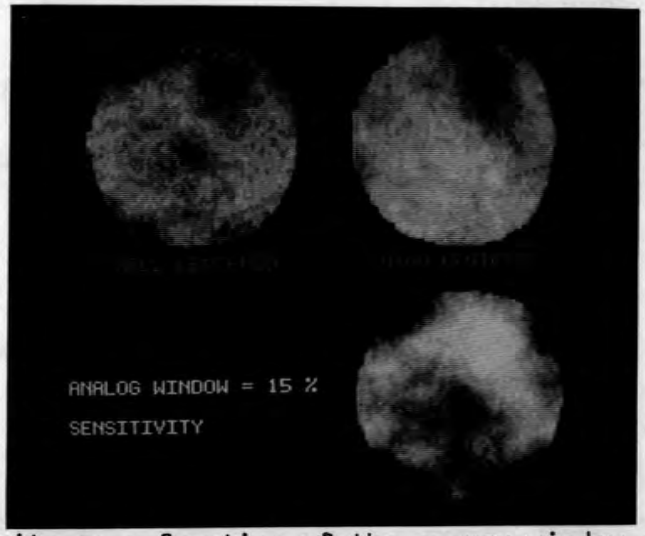


Figure 13: Sensitivity as a function of the energy window setting for a camera with a detuned (lowered) PM. Upper left, the sensitivity matrix for a 15% well centered window; Upper right, for a high (+4%) centered window; below for a low centered window.

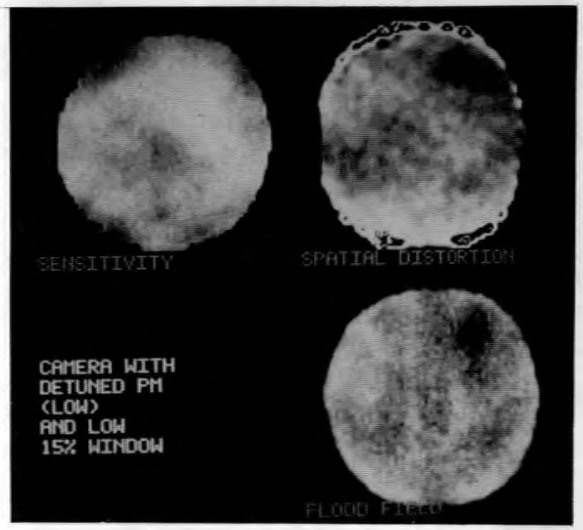


Figure 14 :

Sensitivity (upper left), spatial distortion (upper right) and flood field (below) matrixes for a camera with a detuned (lowered) PM and a low centered 15% window.

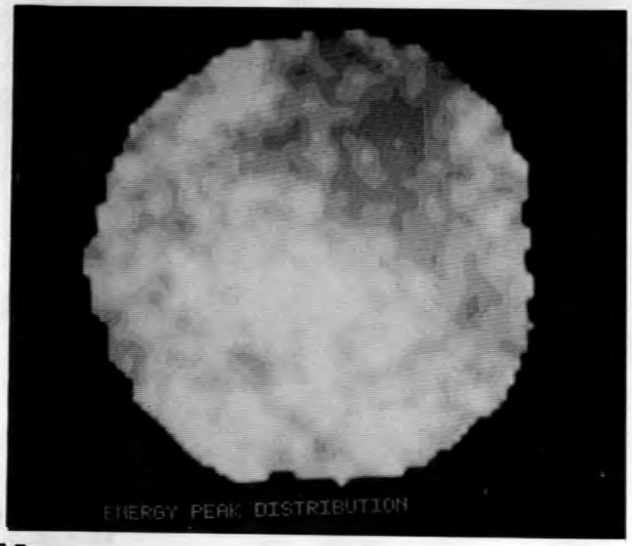


Figure 15 :

Energy photopeak matrix for a detuned camera, to compare with figure 8.

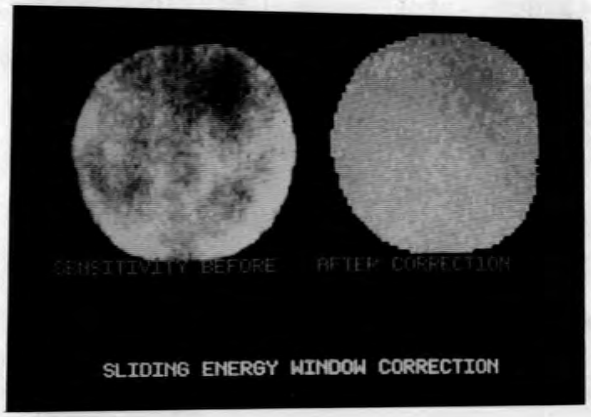


Figure 16 :

Effect of the sliding energy window correction technique on the sensitivity distribution for a camera with a detuned (low) PM. Left, the sensitivity matrix before correction; right, after correction with a constant window width of 10 keV.

constant, especially in an in vivo situation where the width of the window is even more critical due to the varying amount of spatial scatter.

Fig. 16 shows a sensitivity distribution before and after correction by the sliding energy window technique. The analog energy window was 25% (20keV for Hg197). The overall amplitude of sensitivity variation is improved from ± 3.5% to ± 2%, and the hole due to the detuned PM partially eliminated. It is not clearly understood why the technique does not completely compensate for the detuned PM, even when taking into account the fact that the digital relative window width should be variable.

DISCUSSION

The sliding energy window technique clearly improves the variations in sensitivity. For quantitative studies, unless the camera has been tuned such that the energy spectrum (as observed with a multichannel analyser) is constant over the field of view, then this correction technique should be very useful in clinical studies, using a continuous matching of the window width to the observed photopeak as has now been commercially implemented [9].

The physical causes of spatial distortion are not perfectly understood. If it is recalled that a hole in the spatial distortion matrix corresponds to an area of decreased bin size, this might be due to the fact that since the gain of the PM has been lowered, the counts occurring in this particular area are 'repelled' to the nearby PMs. On the other hand, when the gain of a PM is raised, it can be imagined that the counts are attracted to that area. Note that this effect seems to be independent of apparent window setting. In particular it will be necessary to repeat measurements for various scattering thicknesses, as all the work described here was performed in air. The interrelationship between scatter window width and non-uniformity is complex [3].

Clearly, to investigate this effect completely, it is fundamental to measure the distortion as precisely and repetitively as possible. In the opinion of the authors, this cannot be performed satisfactorily using either conventional grid patterns or lead phantom, with holes spaced in a regular lattice [10]. In particular, neither of these methods eliminated the tail effect.

CORRECTION OF SPATIAL DISTORTION

When the sampling size of the displaced point source was more than 5mm, the results indicate that the spatial distortion cannot be measured with an adequate precision in order to attempt any correction.

With the 3.5 mm sampling which was extensively used, the precision for the spatial distortion measurement was of the order of 0.5%. From this information, a transformation may be derived, as defined above, which places photons back to their true location. Using the five parameters : rotation angle θ, scale in X-a, scale in Y-b, and origin Xo,Yo it is possible to calculate from each (X,Y) position measured, the true location (I,J). Although these values are very dependent on the calibration points (which are difficult to choose), the error involved will be of little importance. There is indeed an error in each of the five parameters, but for θ ; Xo,Yo the error will have no consequence causing a small translation

or rotation while for parameters a and b the error will just be a global change of scale in the transformed image.

Thus, the only real problem being how to "place" the photons in their true location. Error due to truncation related to the variable size of the (X,Y) cells must be avoided. A weighted or randomized redistribution, comparable to a 4 x 4 point filter of the 256 x 256 matrix could solve this problem, but no practical results have yet been obtained. Following this, a correction of the sensitivity variations could be attempted although it is likely to be of little value since the errors involved are small, and the errors involved in the correction may not improve the situation. A full correction involves use of the information of the sensitivity matrix and of the tails of the point source response at every position.

CONCLUSION

The problem of correcting non-uniformities in the gamma camera is not trivial. The work that has been presented here has enabled one major step forward to be made, that of measuring accurately the parameters concerned, rather than extracting them from other indirect observations. A whole series of questions can now be answered. It is possible to observe the variation in spatial distortion as a function of energy, setting up the gamma camera etc ... which has not been possible before. It is also possible to relate precisely the sensitivity distribution with the energy and spatial resolution local variation. From this information, methods can be proposed for the successive elimination of the major causes of non-uniformity. Nevertheless these corrections seem less important, since the quantitative errors are in fact small, than the now more attainable understanding of the physical phenomena involved and its consequences on routine situation behaviour, and its implementation in the design of camera systems.

REFERENCES

1. E. Oberhausen, K.J. Neumann, W.Schiffler, Methods for adjusting the Anger Camera, Proc. Symp. Medical Radioisotope Scintigraphy, I.A.E.A., Vienna 1 (1969) 145.

2. E.M. Smith, On methods for normalising the response of the Anger Camera imaging system, Proc. Symp. Quantitative Organ Visualisation in Nuclear Medicine (Kenny P.J., Smith E.M., eds) Univ. of Miami Press, Coral Gables Fla (1971) 261.

3. T.N. Padikal, A. Ashare, J.G.Kereiakes, Field flood uniformity correction: benefits or pitfalls ? J. Nucl.Med. 17 (1976) 653.

4. F.Erbsmann, J. Paternot, A. Piepsz, Sensibilité et Homogénéité, Proc. Symp. La Fixation Rénale du Mercure et du DMSA (Raynaud C. ed) C.E.A., Orsay, France (1976) 28.

5. A. Todd-Pokropek, F. Erbsmann, F. Soussaline, The non-uniformity of imaging devices and its impact in quantitative studies, Proc. Int. Symp. on Medical Radionuclide Imaging, Los Angeles, 1976.

6. B.R. Pullan, Personnal Communication.

7. V.A. Brookman, Spatial distortions in gamma camera images related to counting rate, window width, collimator and collimator orientation (Abstract), J. Nucl. Med. 14 (1973) 383.

8. P. George, C. Raynaud, F. Soussaline, Correction automatique de la dépendance en position de l'énergie; Proc. Symp. La Fixation Rénale du mercure et du DMSA, (Raynaud C., ed.) C.E.A., Orsay France (1976) 45.

9. Picker Corporation, Commercial Exhibit, SNM annual conference (1977).

10. G. Muehllehner, Radiation Imaging Device, U.S.Patent, 3, 745, 345 (July 10,73).

ONLINE DIGITAL METHODS FOR CORRECTION OF SPATIAL AND ENERGY DEPENDENT DISTORTION OF ANGER CAMERA IMAGES

BY

Leonard Shabason, Dennis Kirch, Michael LeFree and William Hendee

The University of Colorado Medical Center
and The Denver VA Hospital

ABSTRACT

The present method of correcting spatial nonuniformities in scintillation camera images consists of taking full-field flood data to obtain a matrix of correction factors. These correction factors are used as either regional multiplying sensitivity factors or as criteria for selectively eliminating (skimming) counts from the hotter regions of the image. The only true justification for changing the number of counts on a regional basis would be the existence of spatial variations in the sensitivity across the face of the camera. Three simple objective measurements serve to point out the possible fallacies of this approach and indicate alternative approaches to uniformity correction. This paper is directed toward an understanding of the nature of Anger camera nonuniformity and suggests a possible online digital approach for correcting scintillation image data.

Supported by HEW Grant 5-732-CA09073, DHEW and Veterans Administration Research funds.

ONLINE DIGITAL METHODS FOR CORRECTION OF SPATIAL AND ENERGY DEPENDENT DISTORTION OF ANGER CAMERA IMAGES

INTRODUCTION

The premise of all nuclear medicine studies is that regions of relative increased or decreased activity observed in scintillation camera images have some meaningfull clinical interpretation. For this reason, the presence of spatial nonuniformities (SNu) may compromise the clinical value of scintigraphic image data interpretation. Because of the importance of this performance characteristic, there is a great deal of interest in uniformity correction techniques. Presently, the most commonly used uniformity correction method consists of obtaining a digital matrix of correction factors from a full-field flood. These data are then used to compute multiplying factors or as criteria for selectively rejecting (skimming) an appropriate percentage counts in the final image. Irrespective of the method of implementation, the fact remains that counts are being added or subtracted from portions of the image which assumes that there are sensitivity variations across the camera face. This assumption was tested by measuring variations in the response of several cameras with a collimated point source and comparing these results to spatial nonuniformities obtained from full-field floods. In all cases studied the variation in the point source response was found to be on the order of $\pm$ 2-3 percent, which is considerably less variation than observed when compared to profile slices made through flood image data. Furthermore, there is little spatial correlation between the variation in the point source response along a diameter of an Anger Camera as compared with the corresponding profile slice through a full-field flood on either an integral or differential basis. We therefore conclude that spatial variations in sensitivity across the camera detector make a minor contribution to overall camera nonuniformity.

Further tests have led to the realization that the major source of spatial nonuniformities arises from spatial distortions in the object to image transformation. Earlier theoretical models of scintillation cameras have shown that Anger cameras are susceptible to spatial distortions as physical variables such as light-pipe thickness or the positional calculating circuitry are altered. This effect was studied by imaging an orthogonal hole phantom. A strong inverse correlation was found between the area enclosed by any four holes as compared to the average count density over that region. That is, hole spacing appeared to increase or decrease in regions of lower or higher count density respectively in the full-field flood images. (2)

Possible correction techniques would be based on the calculation of a matrix of positional shift factors. This approach would require that the camera be made spatially equisensitive as a prerequisite to the calculation of shift factors from analysis of gradients in the flood data. These shift factors would then be implemented in an online microprocessor device which would interpolatively compute the shift factors and correct the analog (x,y) coordinates prior to unblanking. The speed and architecture of presently available microprocessors would permit this to be accomplished without increasing the camera dead-time.

EXPERIMENTAL RESULTS

Point Source Response Experiment A simple objective test points out the possible fallacies of the flood uniformity correction scheme based on sensitivity modification. If SNu are actually due to sensitivity variations then the observed count rate obtained by shifting a collimated point source over the face of a scintillation camera should mirror the SNu observed in a full-field flood. A collimated point source of ^{99m}Tc was positioned across a diameter of an Anger camera (collimator removed) at ¼" intervals with counts collected for a fixed time interval. The point-source sensitivity as a function of distance was compared to the SNu of a profile slice through the full-field flood image from a corresponding region of the camera as shown in Figure 1. Since the objective of the experiment was to investigate the operating characteristics of the camera as it is used clinically, no special tuning procedures were performed prior to the measurements. The outer edge of the camera detector was not included in the measurements to eliminate the influence of edge packing.

In all cases studied, the overall variation in the point source sensitivity was found to be approximately ± 2-3 percent, much less than observed variations in the flood profile slice. Furthermore, the point source response does not show the local hot and cold areas at the same relative positions observed in the flood profile slice. We are led to the conclusion that there is little variation in sensitivity across the camera face.

Effect of Spatially Variant Energy Peak A number of factors contribute to variations in the point source response fluctuation such as variations in the NaI (Tl) crystal sensitivity, variations in optical coupling efficiency or spatially variant energy peaks. It is not possible to definatively measure the contribution of most of these variables in instruments residing in a clinical environment. One variable that may be measured quantitatively is the effect of a spatially variant energy peak whose effect may be seen in Figure 2. As the point source was shifted across the camera detector, a multichannel analyzer was used to record energy spectra as a function of position. The energy scale was established by flooding the camera with ^{99m}Tc and ^{57}Co sources. The results of the shift of the photopeak as a function of position are shown in Figure 2.

In all cases studied the largest overall photopeak shift observed was 5 KeV. A simple calculation based on a Gaussian photopeak approximation demonstrates that for the cameras studied observed shifts of this magnitude have a small effect on count rate. If one assumes a Gaussian shape to the photopeak with 15 percent full-width-half-maximum (FWHM) energy resolution at 140 KeV, the consequences of a 5 KeV shift with a 25 percent window leads to about a 2 percent change in area under the peak. This indicates that a spatially variant energy window of fixed width would do little to directly solve the SNu problem. It would be possible, however, to adjust the lower level of a spatially variant window to achieve a uniform flood condition. While this seems a more intelligent approach than randomly dropping counts it offers little in the way of a genuine solution to the SNu problem.

Effects of Spatial Distortions on Count Density From the above discussion, we are led to the conclusion that the variation in camera sensitivity, whatever its origin, is not the main source of SNu. The major contribution to the SNu of scintillation cameras originates from distortions in the object to image transformation. Relatively small spatial distortions may lead to significant local variations in count density. This conclusion comes as no surprise in light of the results of Svedberg (1) and Cradduck (2) which presented theoretical models of Anger Camera performance. Their calculation showed that images

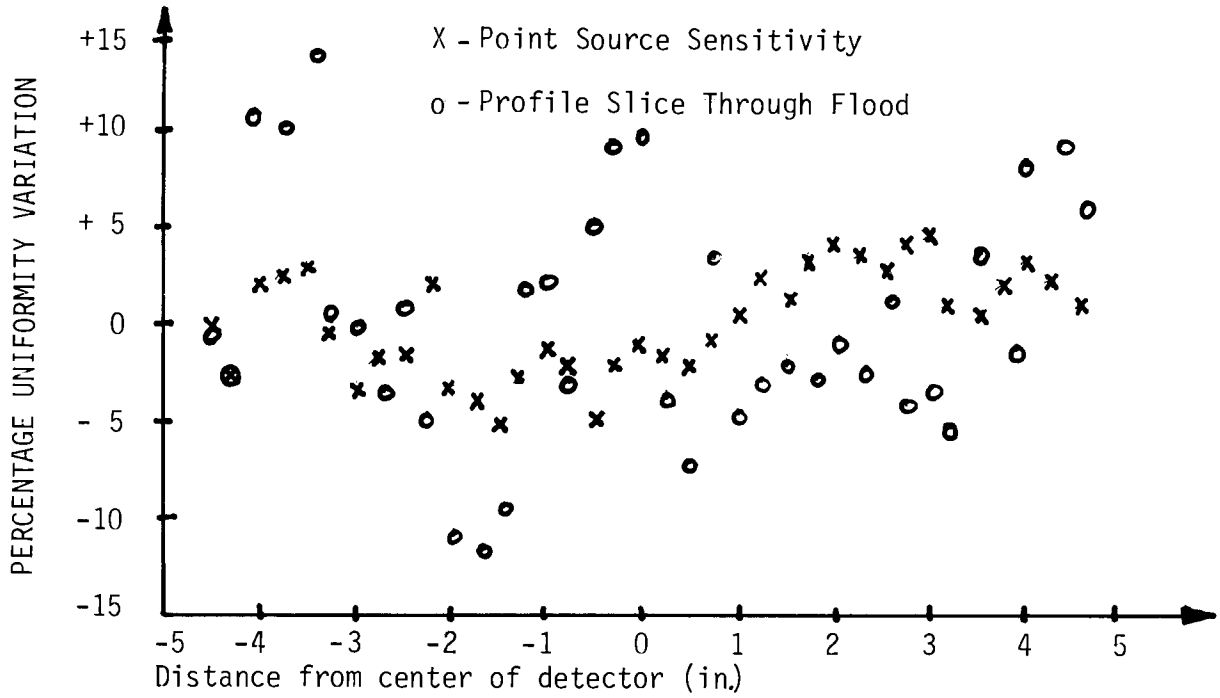


FIGURE 1. Plot of point source sensitivity measurements and a profile slice through the corresponding diametric region of the Anger camera detector. Note that there is little spatial correlation between these two measurements and the range of the deviation in the flood profile slice is much greater (±10%) than the deviation in the point source sensitivity (±3%).

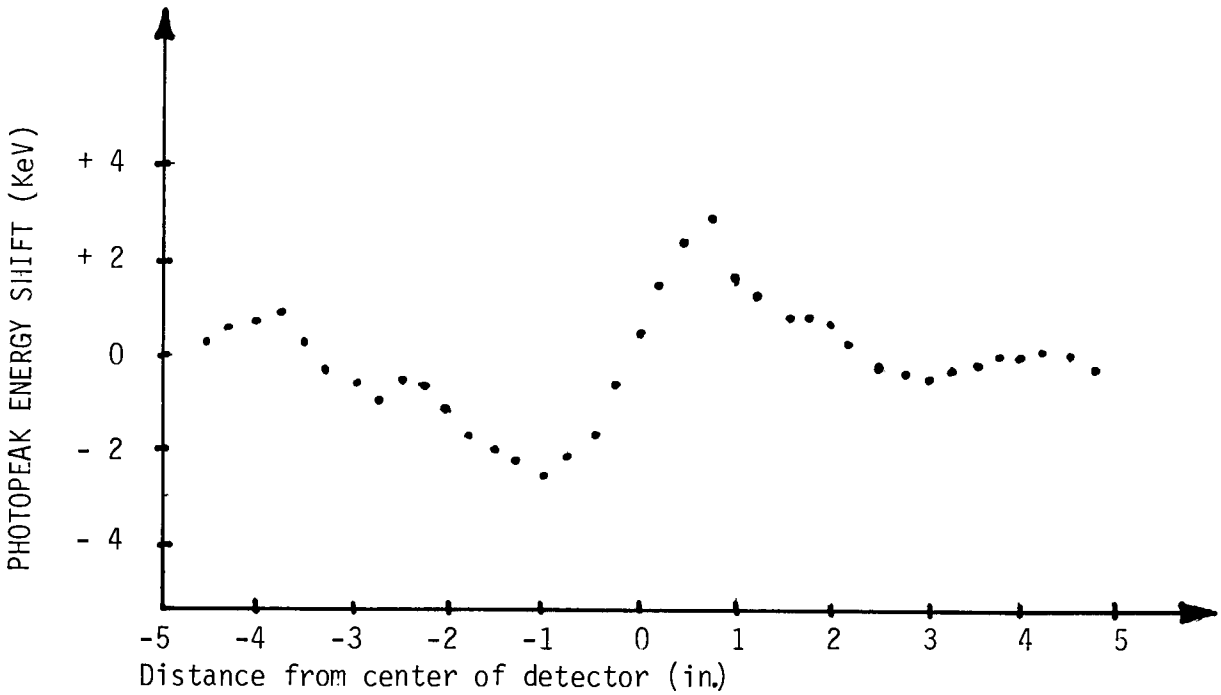


FIGURE 2. Plot showing shifts in the photopeak location of the energy spectra taken at the locations corresponding to the point source measurements in Figure 1. Note that theoretical estimation using a Gaussian approximation and the photopeak shifts shown above predict point source sensitivity variations of the magnitude shown in Figure 1.

are susceptible to distortions when physical parameters such as light guide thickness or the combinatorial algebra used to calculate event coordinates were varied. Experimentally these effects have been reported by Brookman (3).

In order to study this effect, images obtained from an orthogonal hole phantom (OHP) consisting of ¼" diameter holes with ½" separation center to center were compared to full-field flood images. Figure 3 shows an expanded 9 cm square region of an Anger camera image where the crosses identify the position of the maximum count density associated with each hole of the OHP image superimposed on the isocontour lines generated from a full-field flood image. If the area described by each group at four dots is plotted against the average count density in the inscribed region, there is a rough correlation between decreased area and increased count density as is shown in Figure 5. The inexactness of the relationship is due to the superposition of the + 2-3 percent sensitivity variation and the coarseness of the sampling by the ½" OHP. In general, hole images appear to be pulled together or spread apart for high or low count density regions, respectively. The two contours shown in Figure 4 inscribe the highest count density region of the image. Upon qualitative visual examination the general behaviour of the hole locations shows them to be shifted inward toward the higher count density areas and away from areas of low count density.

PROPOSED UNIFORMITY CORRECTION METHODS

It is well known that the signal strength produced at the photomultiplier outputs of the Anger camera is highly dependent on the distance of the scintillation interaction from the center of a photomultiplier tube. This characteristic of the Anger camera is of major importance as the light pipe coupling between the crystal and the photomultiplier is reduced in thickness. The result is shifting of the photopeak several KeV and erroneous displacement of the (x,y) coordinates from their true locations on a spatially varying basis. A variety of electro-optical methods are presently being employed to compensate for these distortions. More recently, a microprocessor based digital device has been incorporated into one of the commercially available cameras (Ohio Nuclear Sigma series) which corrects the Anger camera uniformity by skipping an appropriate percentage of the events in a given segment of the detector as necessary to produce a uniform flood. Todd-Pokropek (4) and others have shown that this type of process can potentially distort dynamic time-activity curves. He has proposed a spatially variant energy window which helps the uniformity somewhat and does not distort dynamic data.

We have performed extensive measurements on two state-of-the-art cameras (O.N. Sigma 410 and Picker 4/15) and conclude that spatial distortion of the (x,y) coordinates is by far the major contributor to Anger camera nonuniformity. It is also true that state of the art microprocessors have achieved speeds which would make it feasible to store appropriate correction factors and relocate the event-by-event coordinates to locations that are more nearly correct. This data also indicates that these cameras are extremely stable and that, given a method for measuring and storing the energy and spatial distortion characteristics, it becomes feasible to make corrections which would compensate for these inaccuracies on a regional basis. The data distortions are fairly complicated and nondeterministic functions which vary over the detector field of view and we conclude that digital storage and online correction of the data is the only reasonable approach to the problem. Offline methods for this kind of problem are operationally awkward and require digital memory sufficient to hold an entire static image.

FIGURE 3. Superimposed isocountour plot from a 9 cm square portion of a full-field flood with the imaged locations of the holes in an orthogonal hole phantom. The hole locations are shown by cross hair symbol and the intensity of the isocountours is proportional to the count density level represented.

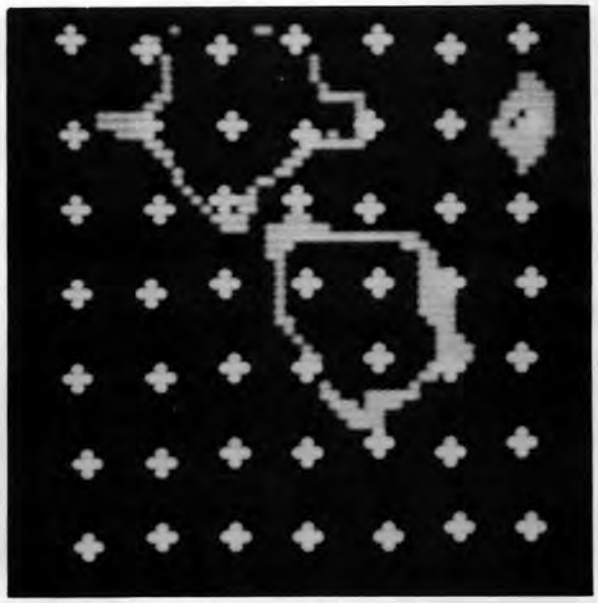


FIGURE 4. A single mean isocountour from the data of Figure 3 shown with the locations of the holes of the orthogonal hole phantom. Note that the grid holes are displaced toward the hot areas of the flood in a direction perpendicular to the isocountour (i.e., along the isoclines). This finding leads to the model in which the negative gradients in the full-field flood data define the direction and magnitude of the dot displacements.

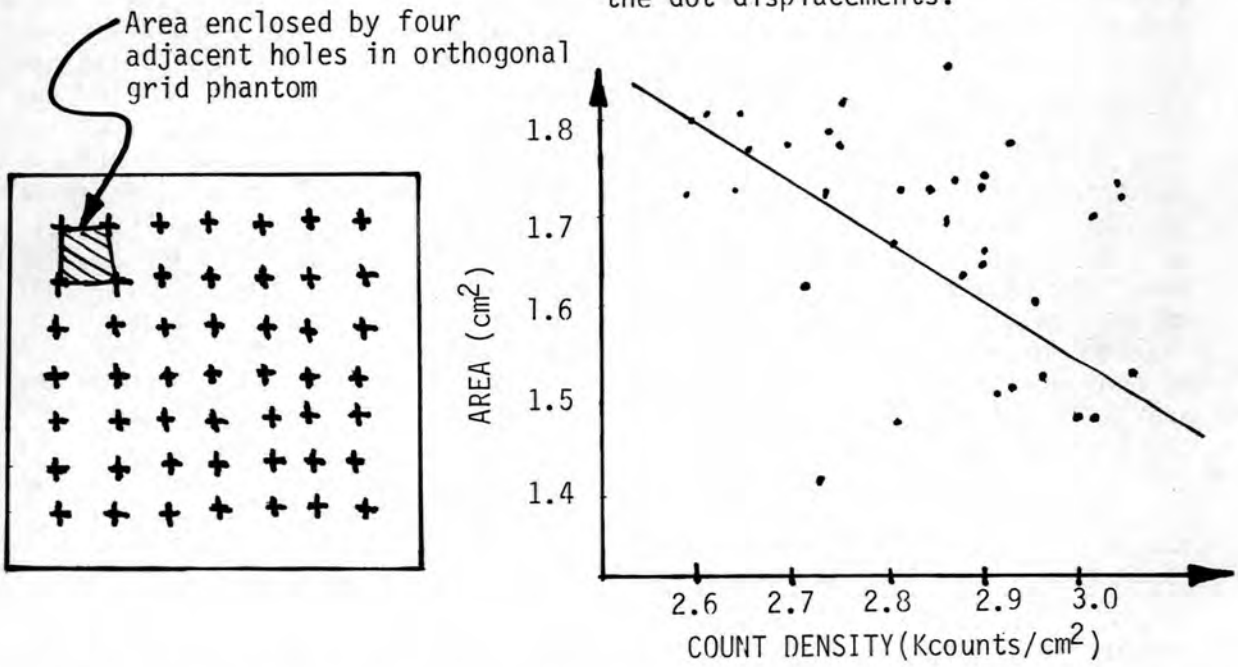


FIGURE 5. Plot showing the rough inverse correlation which exists between the area enclosed by groups of four adjacent orthogonal grid holes and the count density over the corresponding region. This result also supports the concept that dots are pulled apart in cold regions and pushed together in hot regions. The imperfection in this relationship is due to the fact that the Anger camera on which these measurements were made is not perfectly equisensitive.

The most demanding aspect of this problem concerns the method for determining the erroneous displacement of the (x,y) coordinates. Muehllehner (5) has proposed that this be done using a lead plate with holes drilled in a rectangular pattern. This is laborious and time consuming and would be difficult to control in a repeatable fashion. We have developed a shifting algorithm based on the simple full-field flood data alone.

The same digital architecture that would perform the dot shifting could also utilize stored energy window values that correspond to the photopeak value for a particular segment of the crystal. This problem becomes tricky when one considers that the photopeak can change its characteristics drastically in the presence of varying amounts of scatter between the source and the collimator. We have determined a scheme for continuously monitoring the energy spectrum and developing a spatially variant energy window which will adapt to the current conditions of scatter, photopeak FWHM and camera drift and for the first time allow asymmetric windows to be used for camera images.

The concept and architecture of this device are embodied in a dual, twelve bit microprocessor system (Intel 3000 series) which utilizes a 4096 word memory. This processor operates at sufficiently high speeds (200 nanosecond clock rates) to allow the corrections and energy discrimination processes to be performed within the 5 microsecond time limit necessary to prevent increase in the overall dead-time of the Anger camera.

THEORETICAL MODEL OF ANGER CAMERA NONUNIFORMITY

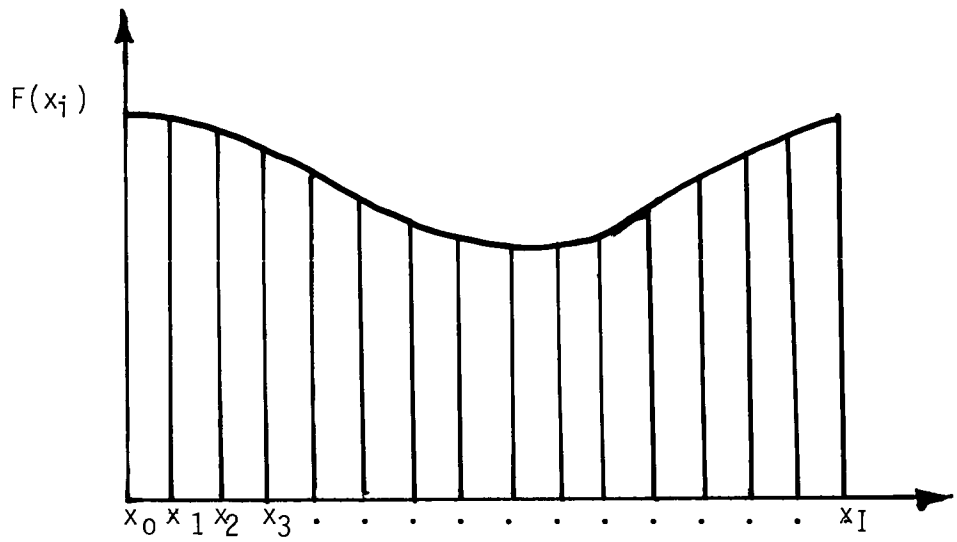
Under the assumption that the camera is equisensitive to a collimated point source placed anywhere on the surface of the detector, we can begin to study the relationship between spatial nonlinearity and full-field flood nonuniformity. At this point one begins to conjour up a number of qualitative analog models. An incorrect analog model is that of a "sand-pile." That is, thinking of hot areas of the field as places where grains of sand have been pushed together and cold area as places where the sand has been pulled apart. The discrepancy present in this model concerns the fact that grains of sand in a pile are able to follow more or less random pathways as they are pushed about to create hot and cold regions. That is, the pathways of any two grains are not constrained to be nonintersecting as the grains are displaced. It is well known and documented by a tremendous amount of published work in which a variety of rectangular grid and bar phantoms have been imaged (3,4) that no matter how drastic the Anger camera nonuniformity becomes, the image of a phantom never crosses, folds or convolutes back upon itself. For this reason we conclude that the sand-pile model for spatial nonlinearity in which the grains of sand are allowed to move along criss-crossing pathways does not fit the Anger camera situation. What we require is a particulate model in which the relative magnitudes of the particle coordinates are preserved throughout the displacements. A more suitable model, therefore, would be that of a rubber sheet with folding of the sheet being disallowed. In mathematical terms, if we move particle #1 from its indistorted position (x_1,y_1) to a new location (x_1',y_1') and a second particle from (x_2,y_2) to a new position (x_2',y_2') then we require that the straight line defining the minimum length pathways associated with these displacements never cross.

The above considerations lead to the development of a relatively clean and simplistic mathematical description of the "rubber-sheet model" and an architecture for implimentation as an online digital correction device. The work that has been done on this problem in the past (4,5) has been based on a point of view that considerably complicates the problem. That is, the previous assumption was that the dot displacements could only be determined by direct

measurement using a precision rectangular grid phantom or a single source stepped around a predetermined grid by some mechanical device. While this approach does not require that the camera pass the equisensitivity requirement, we have found that most state-of-the-art cameras come very close to this assumption to begin with. From a practicle standpoint the direct measurement of Anger camera spatial nonlinearity is a laborious and time-consuming task. Our assumption of equisensitivity and "rubber-sheet model" lead to the simplifying conclusion that all of the information about the dot displacement distortions is contained in the flood data alone and we can eliminate the laborious task of making direct measurements of the spatial nonlinearities.

We begin the development of a theoretical model for correction of spatial nonlinearity with a one-dimensional treatment of the problem. Figure 6 shows a hypothetical plot of a function F(x) which is nonuniform in x. The definition of a mapping from x to x' which will make the function F(x') perfectly uniform is quite simple as shown in Figure 6. The two-dimensional problem is not quite as simple but is handled in essentially the same manner as shown in Figure 7. The same type of integral definition of the mapping transformation from the (x,y) coordinate system to the (x',y') can be written for the two-dimensional problem. The difficulty in applying this transformation to the two-dimensional case lies in the selection of boundary conditions and a reference for the two coordinate systems. It should be noted that uniform mapping can be achieved under a variety of boundary conditions and references. The selection is largely one of procedural convenience and computational efficiency. One possible criterion which can enter into the selection concerns the overall distance which the total transformation must displace the dots to get back to a uniform distribution. When we introduce a minimum displacement distance constraint on the shift transformation, a methodology emerges in which the direction of the shifts coincides with the direction of the negative of the gradient at any given point in the two-dimensional plane and the magnitude of the shift is proportional to the gradient. Note that by this methodology the events occurring at local maxima or minima will not be shifted at all in the mapping process and events occurring at the steapest points in between local maxima and minima will be shifted the furthest. The notion that this method of determining the direction and relative magnitude of the dot shift is a minimum displacement mapping is offered here on the basis of intuitive reasoning alone. The development of rigorous mathematic substantiation awaits the hand of some clever mathematician.

The intuitive basis for the gradient approach to the uniformity mapping relies very strongly on the "rubber sheet" model discussed earlier. It should be noted that other analogous models are equally acceptable. Distortion of a uniform sheet of charged particles by localized positive and negative poles of various strengths creates a similar situation. The charged particles will be drawn toward the poles of opposite charge until the repulsive forces of the particles themselves balance the attraction of the pole and so called "hot spots" will result. A similar but opposite phenomenon will cause particles to be repelled away from the poles of like charge until the repulsive forces of the particles clustered in the hotter regions are balanced. This model also features a constraint on the pathways of particle displacement to be nonintersecting. It should be noted that the gradient method of developing the shift displacement for the uniformity mapping satisfies this nonintersection constraint. While we are certain that the number of analog models which accurately mimic the nonuniform placement of gamma camera events is probably infinite we are somewhat frustrated by a lack of primary knowledge about Anger camera distortions which ties these models directly to some aspect(s) of Anger camera design. As mentioned earlier, such information is difficult to obtain when the camera must be treated as a whole. We were handicapped by not being able to isolate and study the various parts of the camera individually. Still we

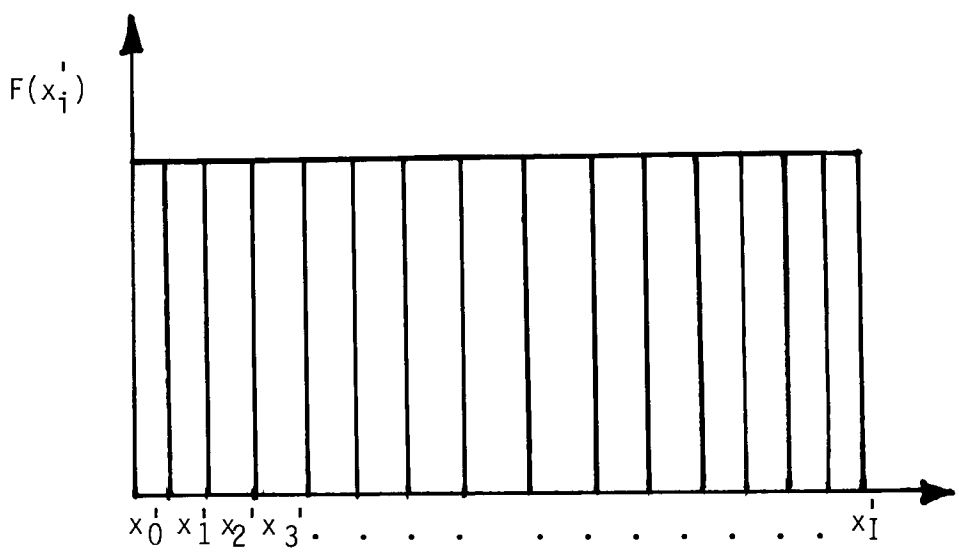


$$T = \int_{x_0}^{x_I} F(x_i)\, dx_i = \text{Total Counts}$$

I = Number of samples

<u>Objective</u>: Achieve mapping of $F(x_i)$ into $F(x_i)$ such that

$$\int_{x_i'}^{x_{i+1}'} F(x_i)dx_i = \frac{T}{I} \quad \text{for all } 0 \leq i \leq I.$$



boundary conditions: $x_0 = x_0'$

$x_1 = x_I'$

FIGURE 6. One-dimensional "rubber-band" model for mapping a nonuniform count density function of x back to a uniform function of x'. Note that the greatest displacement between the nonuniform and the uniform coordinate system (i.e., x-x') occurs where the derivative of the nonuniform function, F'(x), is maximum and no displacement occurs where the derivative is zero.

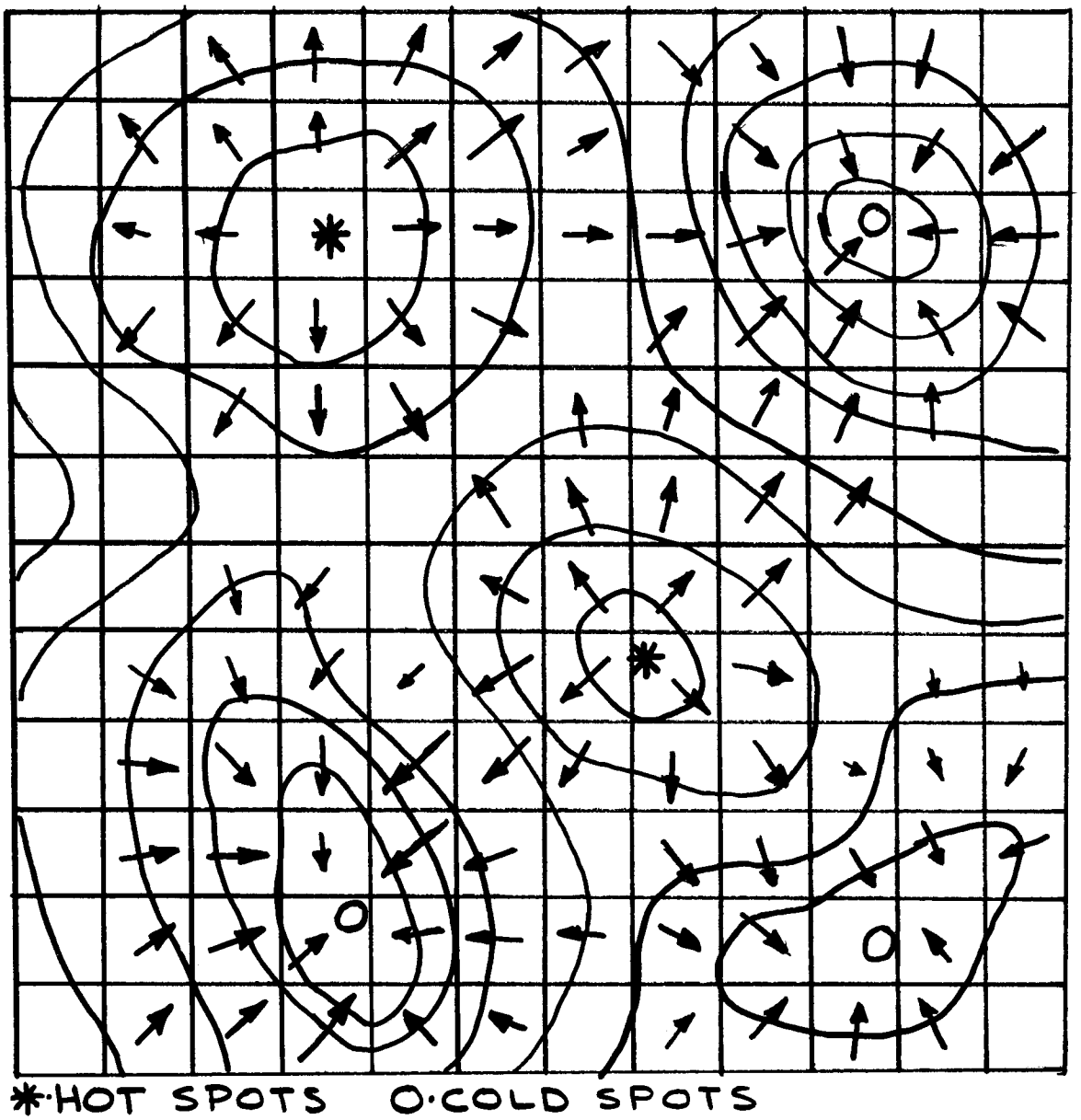


FIGURE 7. Two-dimensional "rubber-sheet" model for mapping a nonuniform count density function of (x,y) back to a uniform function of (x',y'). Local maxima associated with hot spots are indicated by an "*" and local minima by an "o". Isocountours surround the local maxima and minima. The grid represents the pixel boundaries in digital memory where the local shift vectors are stored. Shifts are indicated by arrows which are parallel to local isoclines and are proportional to the negative of the gradient in the full-field flood at that location. In our proposed online digital dot-shifter (Figure 8) the four nearest shift vectors to an event will be used to interpolatively compute the shift for a given event.

feel that the intuitive concepts which our three-part experiment has raised are valid and in fundamental agreement with the "rubber sheet" model which we propose. The question of how one might conceive a uniformity correction shceme based on these ideas which can be implemented on a basis which does not violate the integrity of the Anger camera as it now exists is a whole subject in itself.

HARDWARE IMPLEMENTATION FOR UNIFORMITY MAPPING

The fact that dot skimmers based on field-flood data can operate with sufficient speed to make the correction without increasing the Anger camera dead-time raises the possibility that a dot shifter based on the uniformity mapping ideas presented here could also be implemented on a real-time basis. The architecture of a digital microprocessor device based on this notion is shown in the block diagram of Figure 8. Note that the assumption of camera equisensitivity has been somewhat assured by the introduction of a spatially variant energy window. This is to eliminate the small but measurable variation in sensitivity due to pulse height variations that occur over the surface of the detector. At this point we feel that the full-field flood data alone will contain the information necessary to make the uniformity mapping transformation as previously described. Two very practical questions now arise concerning the necessary statistics and analog-to-digital converter linearity necessary to ensure that the uniformity mapping is accurate. Our calculations indicate that a 64 by 64 flood memory will require that 20 million counts be accumulated to obtain a one percent standard deviation. An even more problematic aspect of this problem concerns the differential linearity of the A/D converters used. In our opinion 8-bit A/D converters which are presently of sufficient quality to produce acceptance image digitization at this time are totally unacceptable for producing the flood data for this application. This is particularly true as we introduce the question of conversion times necessary for real-time operation of such a device. In our opinion we feel that off-line techniques are suitable for feasibility studies only. The operational complications that arise from attempting to do routine image uniformity correction by this method in an off-line modality are horrendous.

SUMMARY AND CONCLUSIONS

This paper presents the concept that spatial nonlinearities (dot displacements) are the major contributor to Anger camera nonuniformity. This is based on the notion that the camera is (or can be made) equisensitive everywhere on the surface of the detector. Upon satisfying the criterion of equisensitivity, we have introduced the simplyfying assumption that the information about the displacement of the dots from their true locations is contained in the flood data itself. This notion is in strong contrast with existing concepts which indicate that precision orthogonal hole phantoms or point sources are necessary determine the dot displacements directly. Using an intuitive model of a rubber sheet we have determined that the dot displacements are in the direction of and proportional to the negative of the two-dimentional gradients in the flood data. This idea leads to a relatively simple online digital architecture for making the uniformity mapping without disturbing the operation of the Anger camera as it now exists. The major technical obstacle which presently complicates the construction of such a device is the lack of A/D converters with sufficient differential linearity and speed to produce data of the necessary quality to make an accuarate mapping possible.

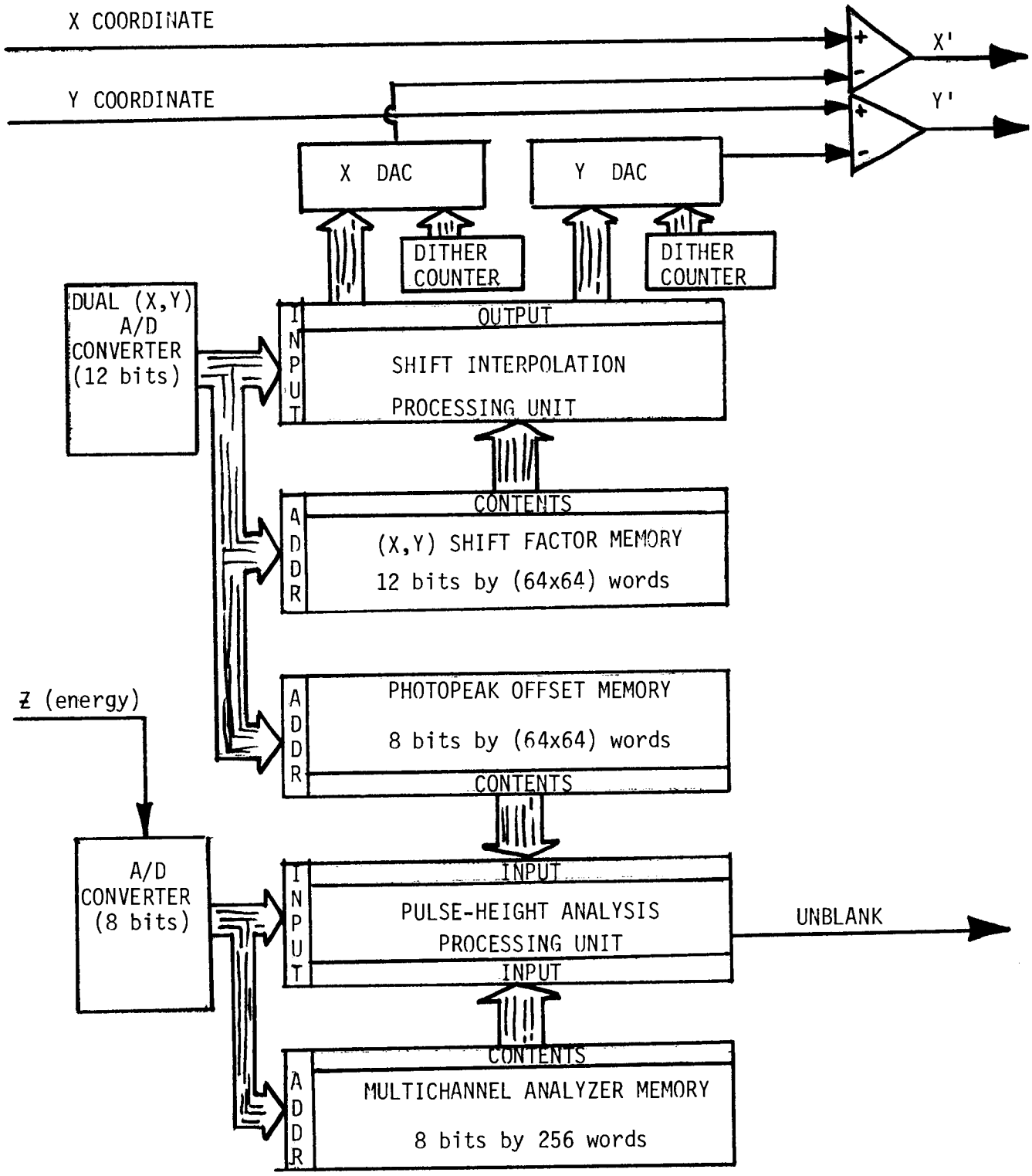


FIGURE 8. Block diagram showing configuration of an online digital device for computing uniformity correction shift vectors. Since the proper function of this device is based on the assumptions that the camera is equisensitive, it is necessary for this device to provide a spatially variant, adaptive energy discrimination capability as shown in the lower-half of the diagram. The upper-half shows the dot shifting portion of the device which looks up the shift vectors nearest an event and interpolates these vectors to determine the shift for a given event. It will probably be necessary to "dither" the shift corrected coordinates to mask out any round-off errors.

REFERENCES

1. J.B. Svedberg: "Image quality of a gamma camera systems", Physics in Med and Biology 13, 597-610, (1966).

2. T.D. Craddock, S.O. Fedoruk: "An experimental determination of the overall spatial resolution of a scintillation camera", Physics in Med and Biology 10, 67-76, (1965).

3. V.A. Brookman: "Spatial distortions in gamma camera images related to counting rate, window width, collimator and collimator oreintations", (abstract), J Nucl Med 14: 383, (1973).

4. A. Todd-Pokropek, F. Erbsmann and F. Soussaline: "The nonuniformity of imaging devices and its impact in quantitative studies", in the Proceedings of the IAEA. Medical Radionuclide Imaging Vol I, IAEA-SM-210, 67-84, (1977).

5. G. Muehllehner: "Radiation imaging device", U.S. Patent No. 3, 745, 345, July 10, 1973.

ALGORITHMS FOR EKG-GATED CARDIAC SCINTIGRAPHY WITH HIGH FRAME RATE AND RAPID TURN-AROUND

Mark L. Moore and Stephen M. Pizer
University of North Carolina
Chapel Hill, NC, U.S.A.

ABSTRACT

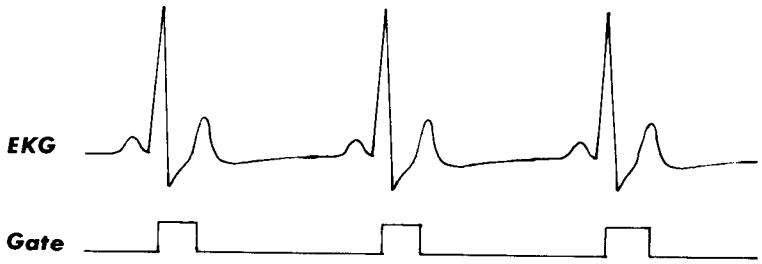
EKG Gated Cardiac Scintigraphy is a useful tool in evaluation of heart function and wall motion abnormalities. The special nature of this problem must be considered before reasonable decisions concerning computer systems to produce these images can be made. This is best done by considering the architecture, implementation and realization of various existing and proposed gated scintigraphy systems. Special emphasis is given to the architecture, i.e. the system as the user sees it, and the implementation, i.e. the algorithm used. A number of algorithms are presented, and the reasons for choosing a particular one provided. The algorithm of choice has been realized on a Digital Equipment Corporation PDP-11/20 and provides 52 64 x 64 16-bit frames per cardiac cycle 1.1 minutes after data collection is complete.

INTRODUCTION TO GATED SCINTIGRAPHIC IMAGING OF THE HEART

Gated scintigraphic imaging of the heart, first developed in 1968 [1], provides a means of imaging the heart's motion. Using the assumption that each beat of the heart is essentially identical to every other beat within any particular individual for a sufficiently short period of time, it sums the images of corresponding intervals of many beats to produce a series of images of appropriately high statistical quality over the period of one typical beat. Based on the above assumption the algorithm operates as follows.

A radiopharmaceutical is introduced into the bloodstream, and is allowed to distribute uniformly throughout the circulating blood. A device called a physiological synchronizer, or gate, takes the EKG as input and identifies occurrences of certain EKG

waveforms in each beat (see figure 1). Each beat is then divided into any number of intervals using the gate signal as a reference (see figure 2). Then, using a scintillation camera scintigraphic data is collected, while monitoring the output of the gate. All counts detected during the same interval with respect to the gating signal are grouped together to form a single image. In this manner, data collected from many beats when the heart is in a particular part of its cycle is used to produce an image of the heart at that part of its cycle.



Output of Physiological Synchronizer

Figure 1

Dividing Beats Into Slices

Figure 2

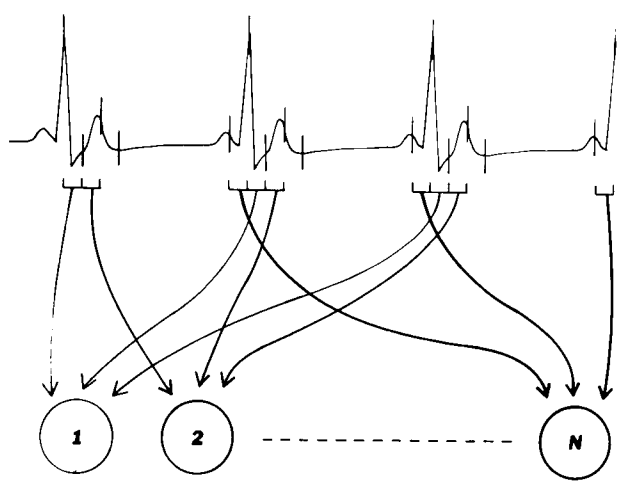
There are many ways to provide these images. One only needs some appropriate type of memory and a way to provide synchronization of data collection with the beating of the heart. In the simplest case, the memory is photographic film. End-systole and end-diastole represent the extremes of the cardiac cycle, the two most interesting parts of the cycle. Using a simple electronic switch, a gate, a timer and a pair of cathode ray tubes, images of these two phases can be obtained [2]. An extension to this method uses a multiformatter. This provides a series of images in which the change between sucessive images is quite small [3]. Other methods use a computer. Systolic and diastolic images can be provided through the use of a pair of buffers in the main memory of the computer. As more frames are desired, the problem becomes more complicated. More buffers are needed and more data must be collected to provide sufficient counts for each image.

ISSUES AFFECTING THE ALGORITHM

The Theoretical Algorithm

Most of the work in gated scintigraphy has used the minicomputer. From a computer standpoint, the problem of producing EKG gated scintigrams of the heart is easy to define: disintigrations detected must be sorted into an ordered set of equivalence classes, one class for each frame in the final image set (see figure 3). Two problems are associated with performing this task.

1) how to define the equivalence classes,
2) how to group the counts which comprise each class together for display.



SORTING INTO EQUIVALENCE CLASSES

Figure 3

The incoming counts form N+1 equivalence classes, where N is the number of frames in the study (time slice divisions of the cardiac cycle). Each count detected during the ith time slice is a member of the ith equivalence class. The order of the N time slices provides the ordering relation on the classes. The N+1th equivalence class is composed of all of those counts which are detected during irregular beats. These are generally discarded, though other alternatives have been suggested.

The sorting problem defined above has the following unique properties:

1) Sorting is into an ordered set of equivalence classes - not into a totally ordered list.
2) The counts do not arrive in a random order, but rather as small groups of counts which belong to the same class.
3) Data can and must be compressed from lists of count coordinates to frames, i.e. square arrays of numbers of counts.

There are also the following practical problems:

4) The minicomputers which are typically used for this application have small main memories and small, slow disk drives, but the problem provides high data rates and a large amount of data.
5) In a clinical environment, the study should be available for review before repeating the acquisition becomes inconvenient, i.e. almost immediately.

Beyond the above computer science problems several difficulties complicate the situation. Defining the time slices is not straightforward. Problems are introduced because the beats are not uniform, so the definition of 1/Nth of the cardiac cycle is difficult. The definition of unacceptable beats is also unclear, as is the choice of a synchronization point (or points) in the cardiac cycle.

Architectural Problems

It is helpful to approach the question of how to provide a reasonable gated scintigraphy system by considering three different ways in which software systems can be viewed. All software systems have an architecture, an implementation and a realization. The architeture is the complete and detailed user specifications. The implementation is the algorithm used, and the realization is the actual coded program. These distinctions are useful because they provide a mechanism to aid in the decisions which must be made when producing any software system. One would like to define the architecture, and then provide an implementation which would allow the realization of the architecture. This frequently does not take place. Gated scintigraphy systems in particular have had their architectures defined by implementation and realization questions. In gated scintigraphy, the key architectural questions are 1) the number and size of the frames, 2) the slicing method used, 3) the wait before the study is available for review, 4) the mechanism provided to deal with irregular beats, and 5) the choice of a synchronization point.

Number and Size of Frames

The ideal number of frames is determined by:

1) the sampling rate required for the blood volume curves for various regions of interest,
2) the required accuracy of frames of the blood distribution at specific times, such as end diastole and end systole,
3) the required smoothness of the motion in motion picture display.

Frames can be added together after the study to increase the number of counts per frame but frames with smaller time increments than those of collection cannot be produced.

The size of the frame matrices, determined by sampling theory from the resolution of the imaging device, is most critical for static images. It also determines the accuracy with which regions of interest can be specified. Note here also pixels can be combined but not broken apart.

The number of bits/pixel is determined by the desired maximum count in a pixel. Too low a volume assures count limited studies. Lower values are acceptable for motion pictures than for static pictures, since the eye does some time averaging.

Slicing Methods

There are serious problems with the assumption that each heart beat is identical. With a normal heart, there are changes in the relationships between the duration of the various parts of the cycle which accompany changes in rate. As the R-R interval lenghtens, the PQRS interval stays the same, the QT interval increases slightly, and the major increase is in the TP interval. In other words, the coefficient of change of duration of the individual subparts of the cardiac cycle is not the same for each subpart, and this does not consider irregular beats which are either ill-formed or have abnormal lengths for the present heart rate.

There are two basic methods of dividing the cardiac cycle into slices. The first is <u>time slicing</u>, which uses slices of fixed, identical duration. The second is <u>phase slicing</u>, which involves measuring the duration of each individual heart beat and using that to calculate the slice duration for each beat so as to provide an equal number of classes for each beat. It is also possible to use time/phase hybrids, in which the more constant PQRST interval is time sliced, and the TP interval is phase sliced (see figure 4). A final approach worth considering is another time/phase hybrid, the predictive method, in which the slice duration for each beat is determined by some estimate of the current heart rate. For irregular beats it is time slicing, for slow changes in rate it is phase slicing.

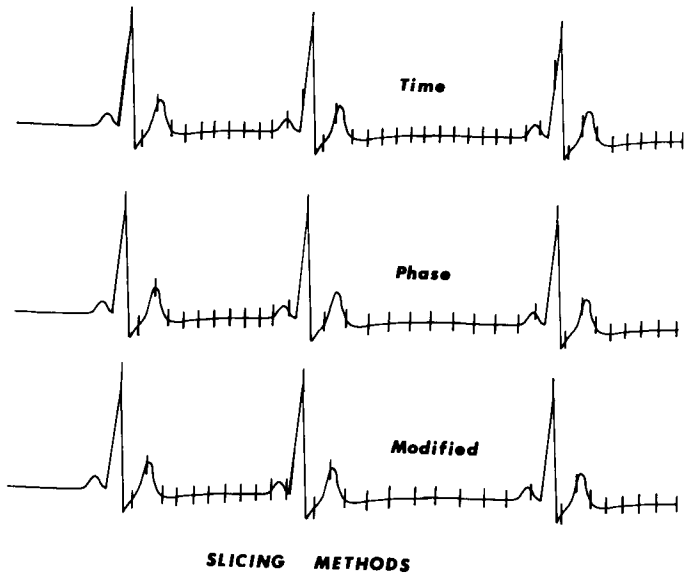


Figure 4

Choice of Synchronization Point

The data collection must be synchronized with the cardiac cycle. Ideally a continuous synchronization would be best, but a suitable mechanism to provide this is yet to be discovered. As a compromise, a point or points in the cardiac cycle are chosen which provide for the best possible synchronization in the parts of the cardiac cycle which are of the greatest interest. The farther from the synchronization point an image is obtained, the more inaccurate the formation of that equivalence class.

The peak of the R wave of the electrocardiogram has enjoyed much popularity as the synchronization point with the cardiac cycle, primarily because of the ease with which it can be detected (an implementation issue). The high rate of change of the first derivative of the EKG at the peak can be identified by analog devices. The R wave occurs near end-diastole, and end-systole is only about 320 milliseconds later, so relatively good images of these points are provided. The possibility of using heart sounds or arterial blood pressure to provide the synchronization point has been discussed, but the peak of the R wave enjoys so many implementation advantages that all existing systems use it. Note that it is possible to begin slicing not at the reference point but at a fixed offset from it.

Rejection of Ill-formed Beats

Data collected during irregular beats, those which are not similar to the norm of the particular patient involved, should not be included in the final picture data because the slicing algorithm breaks down when the assumption of similarity of the beats is violated. The problem is how to identify the irregular beats. As in the synchronization problem, it would be best to monitor the heart continuously to provide very accurate identification of irregular beats. It might be possible to

digitize the entire EKG and perform some type of waveform anaylsis, but the complexity of this seems prohibitive. Assuming the use of a gate which triggers on the R wave, the only information available are the R-R intervals. There is only one possibility for a beat classification system, the method called windowing: those beats in a certain range of durations are accepted, while those which are outside the duration are not accepted. The acceptable range should be a function of the individual patient and his heart rate. An estimate of the current heart rate can be maintained, and the window placed around this estimate.

Insight Into These Problems

At the Hospital La Pitie in Paris France, F.A. Leger, J. Valli and A. Todd-Pokropek [4] have produced a gated scintigraphy system which allows for the investigation of these questions. Their system gates with respect to a variable offset from the R. It allows for both time and phase slicing. A variable window for beat lenghts is provided, and the number of slices can be varied. They conclude that in regular patients time slicing is superior to phase slicing, and that a choice of synchronization point offset from the R wave has no advantage. In irregular patients, time slicing was found to be significantly better than phase slicing and the use of a synchronization point offset backwards from the R wave provides for a higher estimate of the ejection fraction, but no correlation data is presented with respect to non-scintigraphic estimates of the ejection fraction. Windowing is seen as an advantage in irregular patients. Finally, the use of images of less than thirty milliseconds were unsatisifactory. This provides partial answers to many of the questions considered earlier.

These answers and further consideration resulted in the following basic requirements for our system.

1. 40-80 frames per cardiac cycle, to allow accurate identification of the end systolic and end diastolic frames and smooth moving pictures,
2. predictive slicing,
3. 64 x 64 word matrices,
4. study available for review almost immediately, and
5. windowing on beat length.

SOLUTIONS

There are numerous systems which provide solutions to the problem of providing gated cardiac images, and they can be grouped into two functional categories.

1) Those programs which provide 6-16 images during or immediately at the end of the data acquisition.
2) Those programs which provide many images, 10-100, but do not deliver them until hours after the end of the data collection.

Table I provides examples of each category. Note that the figure describes the architecture of the various system. Note the tradeoffs made, especially in terms of the number and size of the frames against the length of time before the images become available for review.

In Core

The work of N.M. Alpert [5] at Massachusetts General Hospital belongs in the first group. Using a PDP-9 (16K of 18 bit words) computer, eight images are collected into main memory as 64 x 64 byte frames. Time slicing is used based on the average heart rate at the start of data collection. There is no transfer of collected counts to a backing store until the completion of the study. Each image has a dedicated buffer in the main memory of the computer into which data is collected. The main memory size limits the number of frames.

A variation of this method was produced by S.L. Bacharach [6], at the National Institutes of Health. By reviewing the problem, he concluded that relatively few images are needed for presentation in movie mode, but that very fine temporal resolution is needed for time activity curves. He provides for the definition of a region of interest (ROI) before the actual start of data collection, based on a few preliminary low count images. Thus, besides 12 32 x 32 word matrices, he creates in core a time-activity curve for the ROI with 10ms temporal resolution.

These methods' in-core data collection with frame organization makes it impossible to reject irregular beats in the frames because the data has already been added to the frames before it could be known that the beat is irregular. The counting rate is not restricted by the algorithm, while the number of images and frequently the counts per image are both severely restricted, both by main memory size.

Offline Reformat

At the National Institutes of Health, M.V. Green [7] has built a system which is an excellent example of the second group. He uses a Hewlett Packard 2100A (22K of 16 bit words) minicomputer to collect list mode data onto magnetic tape. After data collection, the tape is transported to a Digital Equipment Corporation PDP-10 (200K of 36 bit words), where the large main memory facilitates the data reorganization from list mode to frame mode. The PDP-10 then writes time sliced frames of a fixed 10msec duration onto the magnetic tape, which is then returned to the HP 2100A for review. The work mentioned in the last section by Leger is another example of this approach, except reformatting takes place on the minicomputer, with multiple passes of the tape. In both cases the counting rate is limited by the speed of the magnetic tape, which is not a severe restriction. Many frames can be provided, and each can contain many counts. The major problem is the length of time it takes to reformat, which is at least

hours and probably overnight. This approach makes possible extensive data analysis and provides for the ability to reject bad beats, use phase slicing, and manipulate data in other ways. Leger takes advantage of this, while Green does not.

UNC System

Table II provides a summary of the implementation descriptions provided above. Note how the choice of algorithm and the actual target machine have restricted the architecture of the various systems.

For the purposes of comparison, the following is a description of the performance of the system described later in this paper. It provides 52 images per cardiac cycle, each of which is 64 x 64 x 16 bits deep. These images are ready for review about 1.1 minutes after the completion of data collection. It has the disadvantage of limiting the incoming count rate to about 9000 cps and providing only 52 points for the time/activiy curves. It provides windowing for irregular beat rejection and readjusts its estimate of the heart rate continuously. Synchronization at the rise of an externally generated square wave, such as the output of the Brattle gate [2], is provided. The following sections describe the computer system upon which this software system runs and the algorithm used.

Table I

GATED SCINTIGRAPHY ARCHITECTURES

	Alpert	Green	Leger	Bacharach	Moore
Number of Frames	8	100	10-40	12	52
Matrix Fineness	64x64 byte	32x32 byte	64x64 word	32x32 word	64x64 word
Slicing Method	Time	Time	Time or Phase	Time	Predictive
Synchronization	Peak R	Peak R	Offset R	Peak R	Peak R
Availability	Immediate	Hours	Hours	Immediate	2 minutes
Irregular Beats	No action	No action	Sort into Beat Length	Reject in Curves	Reject

Table II

GATED SCINTIGRAPHY IMPLEMENTATIONS

	Alpert	Green	Leger	Bacharach	Moore
Number of Frames	8	100	10-40	12	52
Matrix Fineness	64x64 byte	32x32 byte	64x64 word	32x32 word	64x64 word
Algorithm	In-Core Frame	Offline Reformat	Offline Reformat	In-Core Frame	In Core & on Disk Lists
Computer	PDP-9	HP2100 PDP-10	Varian	HP2100	PDP-11
Main Memory Size	16K 18bits	22K 16bits 200K 36bits	32K 16bits	22K 16bits	28K 16bits

DESIGN OF THE UNC SYSTEM

Gamma-11

The UNC Gated Scintigraphy System had to be implemented on a Digital Equipment Corporation Gamma-11 system. It is located in the Division of Diagnostic Imaging at North Carolina Memorial Hospital. The system consists of a PDP-11/20 with 28K words of sixteen bit memory, complemented by an interface to a gamma camera, a pair of 1,228,000 word removable cartridge disk drives, a general purpose interface, a color display and a Model 33 Teletype. This system is marketed by DEC as an imaging system, with the exception of the color display which was designed and is being built locally [8]. Included in the system is software for the acquisition and anaylsis of scintigrams.

At the heart of the Gamma-11 acquistion is the NC11-A Gamma Camera Interface. Two basic modes of data collection (matrix and list) exist with the NC11-A. In matrix mode operation the data from the camera is recorded in a core matrix. Each element in the matrix represents the number of counts 'seen' in a particular area of the camera face. The camera face is effectively broken down into small areas: the activity in each area is recorded in a specific element of the matrix. In list mode data is recorded in a serial rather than cumulative fashion. Each event is digitized and stored in a core location one greater then the location used for the previous event. In matrix mode the programmer must specify a core location which represents the start of the matrix data, called the offset, and the matrix fineness, either 32 x 32, 64 x 64 or 128 x 128. Also, with the first two matrix finenesses the programmer must specify the amount of storage for each of the matrix elements, either byte or word. Once the interface is started, it will perform non-processor requests (NPR) of the UNIBUS to increment the correct location for each incoming count. In list mode, the programmer must specify a current address, which is the address at which the list is to start. Once enabled, the interface places the coordinates of each incoming count in the list [9].

Further details of the NCMH Gamma-11 system are given in Table III.

Table III

SPECIFICATIONS FOR RK05-AA

MAIN SPECIFICATIONS

Storage medium:	disk cartridge
Capacity/Cartridge:	1,228,800 words
Data transfer speed:	11 microsecond/word
Time for on-half revolution:	20 msec
Disk rotation speed:	1500 RPM

TRACK POSITIONING TIME

One track move:	10 msec
Average:	50 msec
Maximum:	85 msec

DATA ORGANIZATION

Surfaces/drive:	2
Tracks/surface:	200 + 3 spare
Sectors/track:	12
Words/sector:	256
Access with single R/W:	1 to 65,536 words

MAIN MEMORY ORGANIZATION UNDER RT-11

Memory size:	28K
Single Job Monitor:	1.67K
Stack and interrupt vectors:	.5K
Available memory:	25.83K

DISK SPACE AT NCMH

Number of Controllers:	1
Number of Drives:	2
Total available disk blocks:	9600

Algorithms

This section presents the algorithm chosen for implementation, along with a number of the alternatives considered but rejected. For those rejected analysis is limited to the reasons for rejection, and many generous assumptions are made when considering these ill-fated approaches. The discussion assumes that the goal is about 50 frames of 64 x 64 word data which have been time-sliced and windowed.

The information in Table III will be used in the anaylsis of each method. A bastard PL/I description of each method is also provided. See Appendix A for a description of the programming language used.

In-Core Frames

The first method considered is the simplest. Data collection takes place into fixed, in-core frame buffers. It was quickly rejected because of the limitations imposed on the number and size of frames. Only five 64 x 64 word frames can be squeezed into main memory. The use of byte matrices doubles that number, and the use of 32 x 32 byte matrices multiplies it by eight. This approach does not limit the incoming count rate, but it does limit the total number of counts and the size of the frames and does not allow irregular beat rejection.

With the cost of memory dropping as it is, a dedicated hardware system, consisting basically of a large digital memory, could provide the desired images using this approach at a reasonable cost.

```
PERFORM STUDY SETUP;
ON CONDITION (BEAT);
    SET COLLECTION POINTER TO FIRST MATRIX;
    START CLOCK(REPEAT INTERRUPT);
END;
ON CONDITION (CLOCK);
    IF THERE IS ANOTHER MATRIX
    THEN
         SET COLLECTION POINTER TO NEXT MATRIX;
    ELSE
         DO;
         STOP CLOCK;
         STOP COLLECTION;
         END;
END;
ALLOW BEAT INTERRUPTS;
WAIT UNTIL COLLECTION COMPLETE;
PEFORM STUDY CLOSE;
```

Lists to Disk

Collection in list mode and reformatting into frames after data collection is complete was considered next. This approach involves writing the counts to disk in a single list in the order

they are received. Time and beat information must be encoded in the list in order to identify the frame to which each count belongs. The final frames are built after data collection is complete by making multiple passes over the list data or the disk, each time using only the data from the frames which are currently in main memory. The problem with this method is the number of passes which must be made over the list mode data.

```
PERFORM STUDY SETUP;
ON CONDITION (BEAT);
    STOP CLOCK;
    INSERT BEAT MARKER;
    START CLOCK(REPEAT INTERRUPT);
END;
ON CONDITION (CLOCK);
    INSERT TIME MARKER;
END;
ON CONDITION (OVERFLOW);
    WRITE FULL BUFFER TO DISK
    SET COLLECTION POINTER TO EMPTY BUFFER;
END;
DO UNTIL ENOUGH COUNTS;
    ALLOW BEAT INTERRUPTS;
    WAIT UNTIL DISK IS FULL;
    DO J=1 TO NUMFRAMES BY CORE_CAPACTIY;
         READ MATRICES;
         DO I=1 TO NUMBER_OF_BUFFERS;
              READ BUFFER;
              REFORMAT;
         END;
         WRITE MATRICES;
    END;
END;
PERFORM STUDY CLOSE;
```

<u>Anaylsis</u>: -- The goal is to read or write the entire available disk space as fast as possible. Each cylinder of the RK05 contains 6144 words. The fastest way to read or write the entire disk would be to start at the inside or outside cylinder, read the entire cylinder, seek 1 track over, and start reading as soon as the seek is complete. The fastest this could be done is:

1 cylinder rotation + 1 track seek

80 msec + 10 msec = 90 msec

In 90 msec, 6144 words can be written which implies the maximum rate is:

6144 words/90 msec = 68,266 words per second

Assuming about 1800 blocks of disk space is needed for the programs and the frames, 7800 blocks can be used for list. To read or write the list would take at least:

1,996,800/68,266 = 29.2 seconds

The disk is now full of counts, which must be reformatted.
In order to get the best performance, this process must also be disk limited. The actual reformatting which takes place in main memory is sufficiently simple that this is not a problem. To provide the maximum i/o rate, 2 6144 word i/o buffers are needed in core. This leaves room for the program and 2 64 x 64 word buffers. Two frames can be processed in a pass and the counts can be read in at the maximum rate calculated above for collection.

50 frames -> 25 passes
25*29.2=730 seconds to process all frames.

The total time needed to process 1,996,800 counts would be at least 759.2 seconds. If less than 1,996,800 counts were to be used, the virtually unrestricted count rate and the short imaging time could have advantages, but for 1,996,801 counts, there would have to be a reformat before data collection could continue, resulting in an effective count rate of about 2600 cps. This algorithm does not limit the number of frames, or the number of counts per frame, but the collection time increases by a very large constant every time the reformat must take place.

Swapping Frames

The next approach considered was to swap the frames in and out of main memory from the disk, with data collection taking place into the frame buffers in core. Each frame must be in main memory when it is needed. The hope was to use a mechanism similar to the paging approach used in many modern operating systems. Because data collection takes place in an orderly fashion, page faults can be anticipated, and the required frames swapped into main memory before they are needed. This approach is a disaster when the NCMH PDP-11/20 is the target machine because of the limit placed on the number of frames per cardiac cycle, as a result of the slow speed of the RK05 disk.

```
PERFORM STUDY SETUP;
ON CONDITION (BEAT);
    CANCEL PENDING READS;
    WRITE MATRIX BEING COLLECTED INTO;
    SET COLLECTION POINTER TO FIRST MATRIX;
    READ MATRIX 2,3,4;
END;
ON CONDITION (CLOCK);
    IF THERE IS ANOTHER MATRIX
         THEN
              DO;
              SET COLLECTION MATRIX TO NEXT MATRIX;
              WRITE MATRIX JUST COLLECTED;
              END;
         ELSE
              DO;
              STOP CLOCK;
              STOP COLLECTION;
              END;
END;
ON CONDITION (I/O COMPLETE WRITE);
    START READ OF NEXT MATRIX;
END;
ALLOW BEAT INTERRUPTS;
WAIT UNTIL COLLECTION COMPLETE;
PERFORM STUDY CLOSE;
```

Analysis: -- A 64 x 64 word frame is 4096 words long. There is room for five of them in core along with a tiny program. To read each frame would require:

1 track move + 1 aver rotational lat + 1.5 revs read =
10 msec+ 20 msec + 60 msec = 90 msec/frame

= 11.1 Frames/second
= 5.5 Frames/second in and out.

This would allow only 5.5 frames per cardiac cycle in a individual with a heart rate of 60 beats per minute, and only 2.75 in an individual with a rate of 120. The algorithm does not limit the incoming count rate, but it does make windowing impossible.

As an aside, an IBM 2305 Model I (@ approximately $287,000) such as might be used in a paging operating system, has a data rate of 3000 thousand bytes per second. Using that disk, this algorithm could provide a frame rate of at least 120 frames per second.

Lists in Core

Consideration is next given to a method similar to the one above but in which the frame need not be in core whenever counts which belong to it are received. The algorithm proposed involves keeping a list in core of counts for each frame. The image data in frame mode is moved into core from the disk asynchronously with respect to data collection. The appropriate accumulated data is moved from the list into the frame each time it is loaded. Because the lists are always in main memory, there is no need to synchronize the i/o of the frames with the beating of the heart. Also, since the lists should be able to hold the data from a number of beats, the frames can be loaded at a rate less than once every beat.

```
PERFORM STUDY SETUP;
ON CONDITION (BEAT);
    SET COLLECTION POINTER TO FIRST LIST;
    START CLOCK(REPEAT INTERRUPT);
END;
ON CONDITION(CLOCK);
    IF THERE IS ANOTHER LIST
         THEN
              SET COLLECTION POINTER TO NEXT LIST;
         ELSE
              DO;
              STOP CLOCK;
              STOP COLLECTION;
              END;
END;
ALLOW BEAT INTERRUPTS;
DO UNTIL COLLECTION COMPLETE;
    DO I=1 TO NUMFRAMES BY 2;
         READ MATRICES I,I+1;
         REFORMAT APPROPRIATE COUNTS INTO MATRICES;
         WRITE MATRICES I,I+1;
    END;
END;
PERFORM STUDY CLOSE;
```

Analysis: -- Because the frames are being written as fast as possible asynchronous to the collection, they can be carefully placed so that there is no rotational delay after a seek. Using a double buffering technique, two frames are moved for each read or write. On the disk, 2 frames occupy 1.5 cylinders.

The time to read or write 2 frames is:

1 track mov + 2 revs read + 1 track move + 1 rev read =
10 msec + 80 msec + 10 msec + 40 msec = 140 msec/2 frames

In order to write the frames just read, an additional 20 milliseconds would be required to spin around to the start of the first frame.

=160 msec/2 frames
= 6.25 frames per second in and out.

With 50 frames, each list will be emptied every 50/6.25 seconds = 8 seconds.

Assuming the operating system, program and frames require 14K in core, 14K is left for the list buffers. There are 50 lists, one for each frame. Each is 286 counts long, which implies that each list can have a counting rate of:

286/8 = 35.75 cps.

This implies the total counting rate would be:

35.75*50 = 1787.5 cps total.

Here, the algorithm does not limit the frame size or the number of counts, but it severely limits the counting rate, the maximum value of which is a function of the number of frames desired and the main memory size.

Lists in Core - Compressed

The final algorithm presented before the one chosen for implementation is a modification of the above. The in-core lists above should be on the average only half full. If it were possible to use the empty space, this would double the effective count rate. A way to accomplish this would be to have data collection take place into a large circular buffer. As the frames are streamed by as above, the appropriate counts are moved into the frames, and the empty space compressed out of the circular buffer. This compression would actually take place as part of a single pass over the data. Each count in main memory would be considered and either moved to a new location to compress the list, or added to the frame.

```
PERFORM STUDY SETUP;
ON CONDITION (BEAT);
    INSERT NUMBERED BEAT MARKER;
END;
ON CONDITION(TIME);
    INSERT NUMBERED TIME MARKER;
END;
ON CONDITION (OVERFLOW);
    SET COLLECTION POINTER TO START OF LIST;
END;
ALLOW BEAT INTERRUPTS;
DO UNTIL COLLECTION COMPLETE;
    DO I=1 TO NUMFRAMES BY 2;
         READ MATRICES I,I+1;
         REFORMAT APPROPRIATE COUNTS INTO MATRICES;
         COMPRESS LIST;
         WRITE MATRICES I,I+1;
    END;
END;
PERFORM STUDY CLOSE;
```

Anaylsis: -- If it is possible to keep this i/o bound, which is a very generous assumption, the count rate would be at most twice that calculated in 5.4.

1787.5*2=3575 cps.

This is the best method so far. It limits neither the frame size nor the number of counts, but the counting rate restriction is still a problem. Also, the assumption that all the space could be used is generous, and an improvement by less than a factor of 2 should be expected.

Lists on Disk

This approach was chosen for implementation. The algorithm involves keeping a number, k, of lists of counts, with each list containing the data for f frames. Thus the number of frames per cardiac cycle in the study is kf. The lists are stored on disk, and when the disk becomes full, collection is suspended, and the lists are reformatted into frames, compressing the data. Collection then resumes until the lists on disk again become full, when reformatting begins again, now into the already created frames. And so on, until sufficient counts have been collected.

The number of frames per list, f, is determined by considerations of main memory usage during the reformatting process, when list data is compressed to frames. At this time each list is read individually, and the counts from the list are moved to in-core frames. This implies that the number of frames per list is limited by the main memory capacity of the target machine, since all frames for a particular list must be in main memory time during the compression operation.

The number of lists is determined by a more complex process, involving a tradeoff between the desired number of frames in the final study and the desired counting rate, combined with main memory size restrictions during count collection. Fewer lists result in fewer frames but a higher counting rate. The implementor must choose these parameters to provide the architecture specified.

Each list contains the coordinates of disintegrations which are detected during f adjacent time slices of the cardiac cycle, separated by beat and time markers (see figure 5). The data from the first f slices is kept in the first list, the data from the second f slices is kept in the second list, and so forth. The markers, which indicate where within the lists a beat has occurred or a slice has been completed, are introduced into the lists to identify of which frame each count is a member. Each beat marker has a beat number which uniquely identifies the beat. The time information is represented by the list of which a count is an element, combined with the time marker information within a list. The lists are kept on disk, and each list has its own input/output buffers in main memory. Incoming counts are added to the appropriate buffer as they are received. When a buffer fills, it is written to disk. During data collection a list of the identifying numbers of irregular beats is kept. When the disk fills, data collection is stopped temporarily, and the lists of data are reorganized into frames, f at a time, with one pass over the data, ignoring those counts which are part of irregular beats.

There are a number of advantages resulting from keeping the data from f frames together in a list. Since this results in fewer lists, the i/o buffers for each list can be longer. This implies that once a write has started, it will proceed for a longer period of time. More main memory would also provide longer i/o buffers. Since the goal is a disk full of counts, this reduces the number of seeks which must be made.

As little external sorting is done as possible. The data from the f frames within a list can be sorted internally. Even though the counts are not completely sorted when they are moved to the disk, they are sufficiently sorted so as to allow the final sort to take place in a single pass.

When implementing this algorithm on a PDP-11, the problem which must be faced is the long seek time of the RK05 disk drive, typically on the order of 50 milliseconds. Values for k and f fall out of the fixed parameters of the system, such as memory size, the number of words on a cylinder, the seek time mentioned before, and the performance requirements of the problem.

The first parameter to be fixed is f, the number of frames which can be reformatted at one time. Given the choice of 64 x 64 word matrices, main memory limitations determine f as 4. The next decision involves the number of lists to keep. A goal of about 40-80 frames, or about 10-20 lists is the first restriction. The writes are restricted to be a multiple of 256 words by the addressing mechanism of the disk. If in-core buffers of 768 words are used, no write will involve a seek, because 8 buffers will fit exactly on a disk cylinder. Since double buffering is required, each list on the disk requires 1536 words of main memory. Thirteen such buffers fit easily into main memory. Algebra results in a value of 52 for kf, the number of frames in a study. Note that while providing 64 x 64 frames of word data, the algorithm limits the incoming count rate, because the counts must be written to disk.

These assignments for k and f are acceptable. In summary, data collection takes place in the following way. There are 13 lists, each of which has a tail in main memory at collection time (see figures 6,7). Data collection takes place into the areas for each tail. The counts received for slices 1-4 are included in list 1, the counts for slices 5-8 are included in list 2 and so forth (see figure 8). Each list contains the coordinates of the detected disintigrations, beat markers and time markers (see figure 5). When one of the lists fills on the disk, data collection is stopped and the reformat/compress operation takes place. Main memory at reformatting time contains space for 4 frames in additon to 2 buffers to double buffer the counts in core (see figure 9). The steps needed to reformat 4 frames take place 13 times, resulting in 52 frames of data.

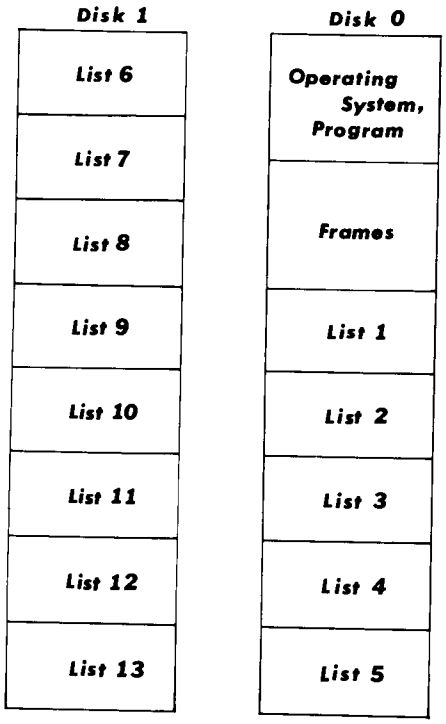
x	y	
	1	–b
1	3	
4	6	
		–t
7	8	
		–t
3	7	
5	5	
6	4	
		–t
1	2	
2	4	
6	8	
	2	–b
7	9	
3	8	
4	2	
		–t
9	1	
		–t
2	2	
6	4	
7	9	
		–t
8	6	
7	9	
1	8	
	3	–b
4	5	
5	4	
		–t
6	9	
5	9	
5	8	
1	1	
		–t
3	3	
5	9	

b - Beat

t - Time

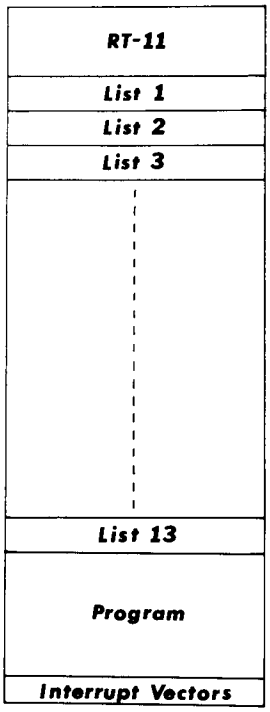
TYPICAL LIST

Figure 5



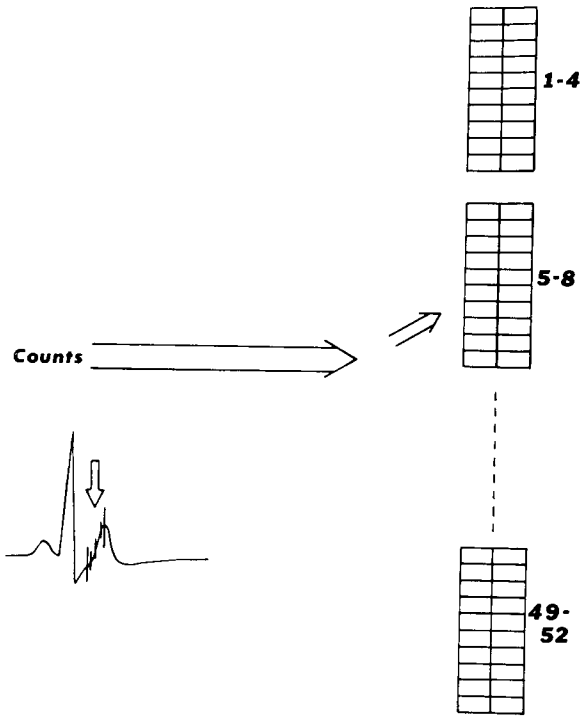
DISK USAGE

Figure 6



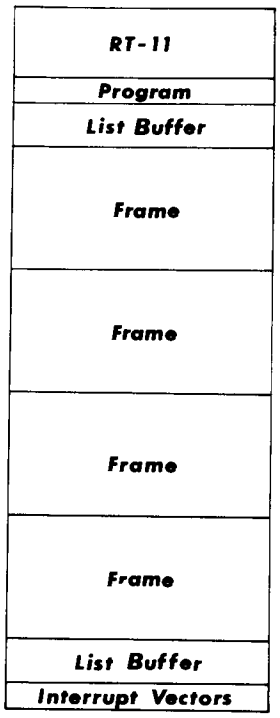
MAIN MEMORY AT COLLECTION TIME

Figure 7



COUNT DISTRIBUTION

Figure 8



MAIN MEMORY AT REFORMAT

Figure 9

<u>Anaylsis</u>: -- With only 13 lists there are 13 (2*768)-word buffers, which can be used for double buffering. Eight fit exactly on a cylinder, so there is never a seek while writing. Each write is:

1 ave seek + 1 ave rot latency + time to write =
50 msec + 20 msec + 10 msec
= 80 msec/buffer
= 12.50 buffers/second
= 9600 counts/second.

These must be reformatted when the disk fills. There are 7800 disk blocks available for lists.

7800 blocks = 1,996,800 counts.

The reformatting will have to take place every time about 2 million counts are collected.

The time to reformat is calculated as follows. This process is i/o bound as well, but since it takes place in an orderly fashion, the timing calculations no longer need consider the average case.

Each list is 600 blocks long. The buffers available in core are each 2048 words long. Each list will require 75 reads, but only every third read will require a seek (an additional rotation). On the disk which contains the frames, an average seek will be required at the start of each list. This will add an additional 250 milliseconds to the reformat.

To read 3 buffers:

1 ave rotational latency + time to read 6144 words
20 msec + 80 msec
= .1 seconds per read
= 2.75 seconds/list
=35.75 seconds to read the lists.

The frames must be read as well:

1 average Seek + aver rotational Latency + time to read
50 msec + 20 msec + 60 msec
= 130 msec/frame = 6.76 seconds to read all frames
13.52 seconds to read and write

Total reformat time = 13.52 + 35.75 = 49.27 seconds

This algorithm has been realized in the NCMH PDP-11/20. The program is broken into three overlays, so as to make as much of main memory available at collection and reformat as possible. The first overlay contains the code to perform setup functions such as the creation of the patient file, the determination of an estimate of the average heart rate and so forth. The second overlay performs data collection, which is interrupt driven, and this overlay consists of little other than the interrupt handlers. The final overlay performs the reformatting. There is a small driver which is always resident, which calls the overlays and maintains global variables.

RESULTS

The program was tested using the turntable phantom described in [3]. Figure 10 shows twelve four-frame summed images from the first forty-eight frames of a typical patient study.

CONCLUSIONS

It has been shown how, for the problem of gated heart imaging with a given set of specifications, the architecture can be specified without concern for the realization, and then a realization can be provided. This realization operates as predicted by analysis. It provides 52 64 x 64 word matrices, data maximum count rate of 9000 counts per second, 1.1 minutes after data collection is complete.

There are many possible additions which could be made to a system using this algorithm. Histograms of beat lengths could be used to provide images which reflect the heart's behavior for specific beat lengths. The data from the first reformat could be displayed as soon as the reformatting is complete. Gating with respect to multiple signals is possible, as is reformatting backwards from the synchronization point. But our purpose here was to describe a robust algorithm which can be used to realize various gated scintigraphy architectures. The count rate/number of frames tradeoff still exists, but this algorithm can be used in EKG gated cardiac scintigraphy systems in which the clinical utility is of prime importance.

Typical Patient Study--Left to Right

Figure 10

STATEMENT	DESCRIPTION
ON CONDITION ( ) on-clause END	The execution of this statement causes no immediate action. Rather, when the condition occurs, the statements which comprise the on-clause are executed. When the on-clause completes execution, program control returns to the point at which it was interupted when the condition was raised.
IF condition THEN then-clause ELSE else-clause	where "condition" is a condition which is either true or false (for example, A < B); "then-clause" is the program section to be executed if the condition is true; and "else-clause" is the program section to be executed if the condition is false. The then-clause and the else-clause are each either a single statement or a group of statements preceded by DO; and closed by END;.
DO index-variable= starting-value TO ending-value BY increment; interior statements END;	This construct has the following effect: First, the index variable is given the starting value in an initialization step. Next, a test is made to see whether the index variable has a value beyond (greater than if the increment is positive, less than if the increment is negative) the ending value. If so, the loop is terminated, and execution continues with the statement following the END statement of the loop. If not, the interior statements of the loop are executed until the END statement of the loop is reached, either in sequence or as the result of a GO TO statement. Then, the increment is added to the value of the index variable, and the sequence described by this paragraph is repeated until the test is satisfied. If the "BY increment" part of the DO statement is omitted, the increment is assumed to be 1.

REFERENCES

[1] C.B. Mullins, D.T. Mason, J. Ross, W.L. Ashburn, "New Approach to the Measurement of Ventricular Volume: Radioisotope Angiocardiography with Technetium 99-M", Circulation 38(Suppl VI):VI-144 (1968).

[2] "Brattle Physiological Synchronizer Model 202 Instruction Manual", Brattle Instrument Corporation, Cambridge, Massachusetts.

[3] E.A. Silverstein, D.A. Turner, E.W. Fordham, A. Chung-Bin, "Cardiac Blood Pool Imaging Over the Complete Cardiac Cycle with a Multiformat Imager", J. Nuclear Medicine 18:159-162 (1977).

[4] F.A. Leger, A.E. Todd-Pokropek, "A Comparison of Software and Hardware Gated Cardiac Studies", J. Nuclear Medicine 18:607 (1977).

[5] N.M. Alpert, K.A. McKusick, G.M. Pohost, R.E. Dinsmore, S.E. Potsaid, "Non-invasive Nuclear Kinecardiography", J. Nuclear Medicine 15:1182-1184 (1974).

[6] S.L. Bacharach, M.V. Green, J.S. Borer, M.A. Douglas, H.G. Ostrow, G.S. Johnston, "A Real-Time System for Multi-Image Gated Cardiac Studies", J. Nuclear Medicine 18:79-84 (1977).

[7] M.V. Green, H.G. Ostrow, M.A. Douglas, R.W. Myers, R.N. Scott, J.J. Bailey, G.S. Johnston, "High Temporal Resolution ECG-Gated Scintigraphic Angiocardiography", J. Nuclear Medicine 16:95-98 (1975).

[8] S.M. Pizer, F.H. Chan, J.M. Fullton, R.E. Johnston, & H.H. Lane, Jr., "A Medical Imaging Display System - Ultrasonography", Paper 96-56, Proceedings SPIE Conference, Application of Optical Instrumentation in Medicine V, (September 1976, Washington, D.C.), pp. 354-360, Volume 96, ISBN 0-89252-123-6, SPIE, Palos Verdes, CA, 1976.

[9] Instruction Manual, NC11A Gamma Camera Front End, DECSPEC-11-ATFADA, Digital Equipment Corporation (1973).

A STOCHASTIC METHOD FOR DETERMINING THE BACKGROUND-TO-LIVER RATIO IN LIVER SCANNING*

B. E. Oppenheim and K. Rezai-Zadeh
Department of Radiology, Indiana University School of Medicine
Indianapolis, Indiana, USA
and Department of Radiology, The University of Chicago
Chicago, Illinois, USA

ABSTRACT

A stochastic method has been developed for determining the ratio of total detected background events to total detected events originating in the liver in scans performed with the scintillation camera. This method has been validated through phantom studies and applied in 45 patient studies. It has led to an understanding of the role of background activity in the centroid-fixation method of motion correction.

INTRODUCTION

In imaging the liver with the scintillation camera and a colloidal radiopharmaceutical, what fraction of the detected events originate within the liver, and what fraction from structures outside the liver? This interesting question has arisen in conjunction with motion correction studies. In the centroid-fixation method of motion correction proposed in digital form by Oppenheim[1] and implemented in analog form by Hoffer et al.[2], motion is corrected by continually identifying the centroid of detected events and shifting these events so that the position of the centroid appears fixed.** The method assumes that all of the radioactivity under observation is moving in unison. If a significant portion is located in stationary structures, however, the position of the centroid does not accurately indicate the position of the liver, and when the detected events are shifted to fix the centroid there will still be some residual liver motion.

We define the background-to-liver ratio r as the ratio of detected events originating in stationary structures to detected events originating in the liver and structures that move with it (mainly the spleen). Now if we knew the value of r we could completely correct for liver motion by increasing the amount of shift required to fix the centroid by a factor 1+r. The value of r can be estimated from a digitized image of the liver and spleen, using a region-of-interest method, in which regions corresponding to the liver and background

*This work was supported in part by the Center for Radiologic Image Research (USPHS Grant GM-18940), Dept. of Radiology, Univ. of Chicago.
**The method originally used the median as the central measure, but the centroid has been found to be preferrable.

activity are identified in the image and the events in these regions are quantitated. This method is time-consuming, however, and assumes that we can accurately estimate the number of background events originating within the liver region, which may not be possible. Therefore we have developed a stochastic method for determining r, based on statistical parameters obtained from images of the moving liver. This method can be automated, and is potentially more accurate than the region-of-interest method.

METHODS

The stochastic method assumes that only two types of radioactivity are present within the image: liver activity, which undergoes uniform translational motion in the longitudinal or y-direction but is otherwise fixed in its spatial distribution, and background activity, which is stationary. Let us say that during some short time interval in which liver motion is negligible, n total events are detected, consisting of n_L liver events and n_B background events. The y-coordinates of the liver and background events are denoted by y_{L_i} $(1 \leq i \leq n_L)$ and y_{B_j} $(1 \leq j \leq n_B)$ respectively. We define $\bar{y}_L$, $s_{y_L}{}^2$, $\bar{y}_B$, $s_{y_B}{}^2$, $\bar{y}$, and $s_y{}^2$ as the means and variances of the y-coordinates of the liver events, the background events, and the total events (the y_{L_i} and y_{B_j} taken together). Note that all of these terms, including the n's, are random variables, but only y_{L_i}, $\bar{y}_L$, $\bar{y}$ and $s_y{}^2$ are dependent upon the position of the liver. We wish to determine r, which is the expected value of n_B/n_L. Now

$$\bar{y} = (n_L\bar{y}_L + n_B\bar{y}_B)/n \quad ,$$

and

$$s_y{}^2 = \left[\sum_{i=1}^{n_L} (y_{L_i} - \bar{y})^2 + \sum_{j=1}^{n_B} (y_{B_j} - \bar{y})^2 \right] /(n - 1) \quad .$$

It follows that

$$s_y{}^2 = \frac{n_L - 1}{n - 1} s_{y_L}{}^2 + \frac{n_B - 1}{n - 1} s_{y_B}{}^2 + \frac{n_L}{n - 1} (\bar{y}_L - \bar{y})^2 + \frac{n_B}{n - 1}(\bar{y}_B - \bar{y})^2 \quad ,$$

also

$$\bar{y}_L - \bar{y} = \frac{n_B}{n_L}(\bar{y} - \bar{y}_B) \quad ,$$

Then

$$s_y{}^2 = \frac{n_L - 1}{n - 1} s_{y_L}{}^2 + \frac{n_B - 1}{n - 1} s_{y_B}{}^2 + \frac{n_B}{n_L}\left(\frac{n}{n - 1}\right)(\bar{y} - \bar{y}_B)^2 \quad .$$

All terms in the last equation are random variables, but only $\bar{y}$ and $s_y{}^2$ are dependent on the position of the liver. We may therefore rewrite this equation in the form:

$$s_y{}^2 = \frac{n_B}{n_L} \bar{y}^2 + k_2\bar{y} + k_3 \quad , \tag{1}$$

where k_2 and k_3 are random variables which are not dependent on liver position. (We assume that n is sufficiently large that n/n-1 can be equated to 1.) Thus $s_y{}^2$ is a quadratic function of $\bar{y}$, and the coefficient of the second degree term is n_B/n_L.

We implement the stochastic method by forming a series of brief images, or

frames, of the moving liver and computing $\bar{y}$ and s_y^2 for each frame. From Eq. 1 it follows that the points $(\bar{y}, s_y^2)$ are distributed about a parabola, the curvature of which is r, the expected value of n_B/n_L. We estimate r by fitting a second-degree equation to the set of points, using a least-squares algorithm[3], and determing the coefficient of the second-degree term. We denote its value as r_{comp}, the computed value of r, as distinguished from r_{true}, the true value of r. The algorithm also estimates the standard error of r_{comp}.

For patient studies the region-of-interest method was also used to evaluate the background-to-liver ratio: The total events were determined within a region-of-interest that included the liver and spleen. The number of these events that were due to background activity was estimated, based on the events per unit area within some representative surrounding region. This estimate was subtracted from the total events in the former region to give the liver events. This value was subtracted from the total events in the image to give the background events. The ratio of background events to liver events found in this manner was denoted as r_{ROI}.

COMPUTER SIMULATION STUDIES

The stochastic method was tested by simulating a series of frames representing brief images of the moving liver. A typical liver count distribution and a typical background count distribution were simulated separately on 128 x 128 grids. Since only y-coordinates are needed to apply the method, longitudinal profiles of these distributions were formed by summing rows. A value for r_{true}, the true background-to-liver ratio, was chosen, and the profiles were appropriately scaled to correspond to this value. The liver profile was shifted longitudinally to 29 different positions, and at each position it was superimposed on the background profile to form a profile of the total count distribution for that position. Three of these profiles are shown in Fig. 1. Each profile simulates the one-dimensional projection of a frame.

A set of 29 profiles was formed for $r_{true} = 0.30$, another for $r_{true} = 1.00$. The corresponding sets of points $(\bar{y}, s_y^2)$ were computed from the profiles and plotted, along with the second-degree equation fitted to these points (Fig. 2). A perfect fit was obtained; furthermore, r_{comp}, the coefficient of the second-degree term, agreed precisely with r_{true}. Next, the random nature of the counting process was simulated by replacing the counts in the profiles by Poisson random variables with expected values appropriately selected to simulate 5,000 counts/frame (Fig. 3). The points $(\bar{y}, s_y^2)$ were now scattered about the fitted curve, and r_{comp} was a somewhat less accurate measure of r_{true}.

To determine the accuracy and reliability of r_{comp} as a measure of r_{true} we set the value of r_{true} at 0.30, the counts/frame at 5,000, and the range of liver motion at 20 rows, and then varied these parameters, one at a time (Figs. 4-6). For each set of values of the parameters 10 trials were run, each consisting of 300 profiles from which 300 points $(\bar{y}, s_y^2)$ were computed. The curve-fitting algorithm[3] yielded a value of r_{comp} and its standard error for each trial. The average of the ten values of r_{comp} and the average of the standard errors were used to give the "unweighted" mean and one standard deviation range of r_{comp}. As shown in Figs. 4-6 the range of error in the value of r_{comp} becomes greater as r_{true} increases or the counts per frame or range of motion decrease; moreover, r_{comp} becomes a slightly biased estimator of r_{true}, as its "unweighted" mean values fall below the corresponding values of r_{true}. This bias was found to be due to the random error in the value of the independent variable $\bar{y}$. In regression analysis it is generally assumed that the independent variable is measured without error; however, if error is present the estimate of parameters is biased[4]. We developed a modified curve-fitting algorithm which weighted the points $(\bar{y}, s_y^2)$ to offset this effect. The resulting "weighted" values of r_{comp} appeared to be unbiased estimators of r_{true} (Figs. 4-6).

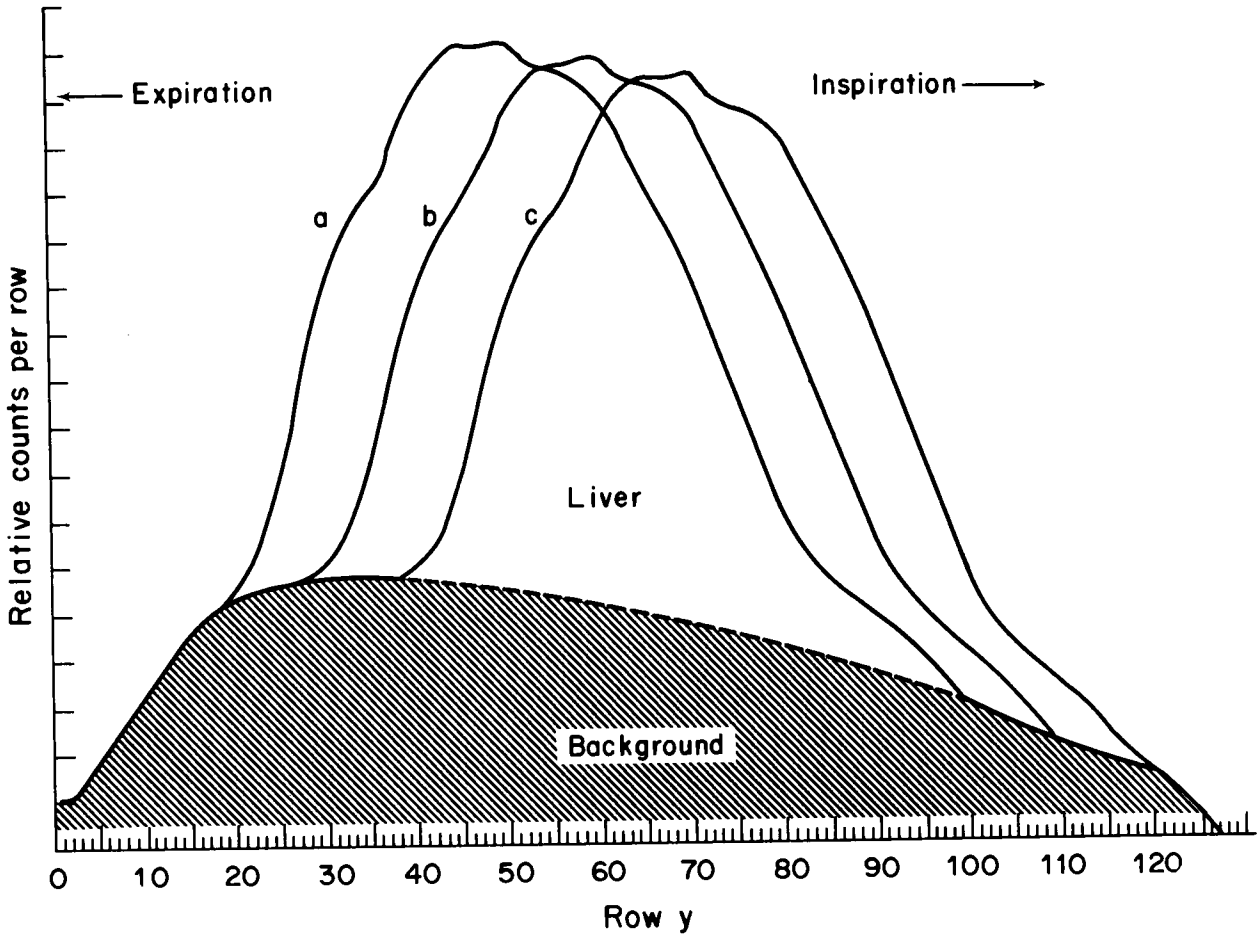


Fig. 1. Longitudinal profiles of total count distribution for three simulated 128 x 128 frames of the moving liver. The liver component moves with respiration; the background component is fixed. a. expiration, b. mid-inspiration, c. full inspiration.

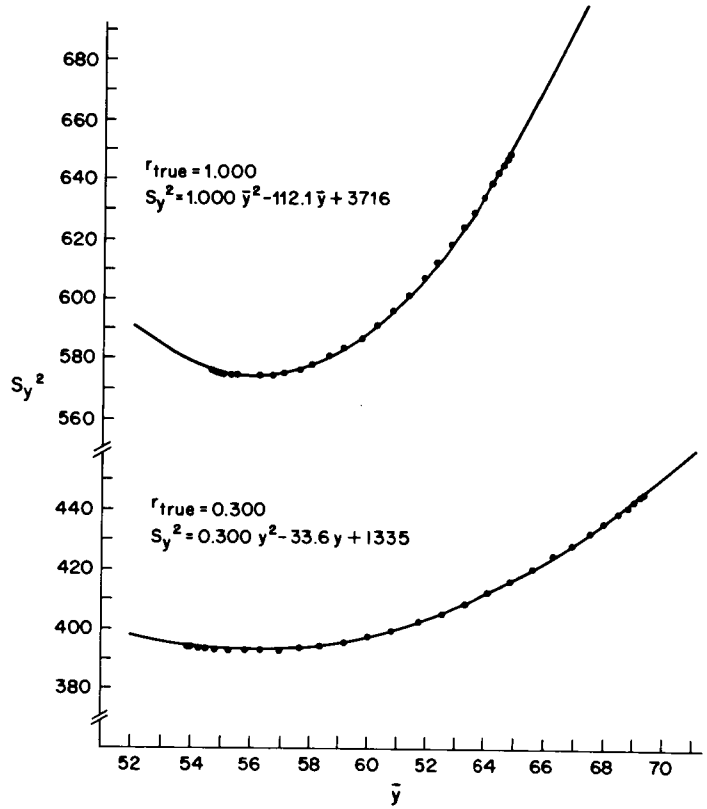


Fig. 2. Plot of points ($\bar{y}$, s_y^2) for frames representing 29 positions of the liver and infinite counts per frame. A parabola was fit to the points by a least squares technique. The coefficient of the second-degree term (r_{comp}) agrees precisely with r_{true}.

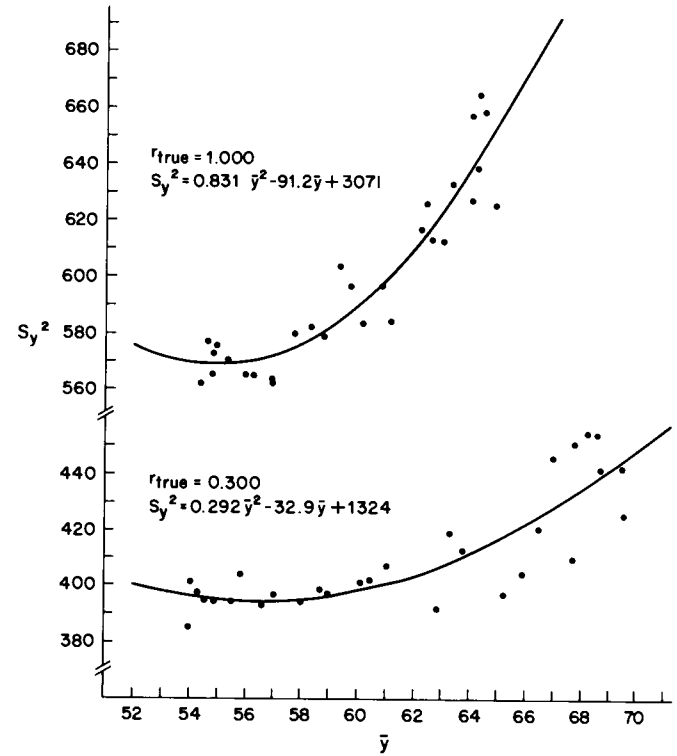


Fig. 3. Same as Fig. 2, except noise was simulated corresponding to 5,000 expected counts per frame. The coefficient of the second-degree term (r_{comp}) agrees less well with r_{true}.

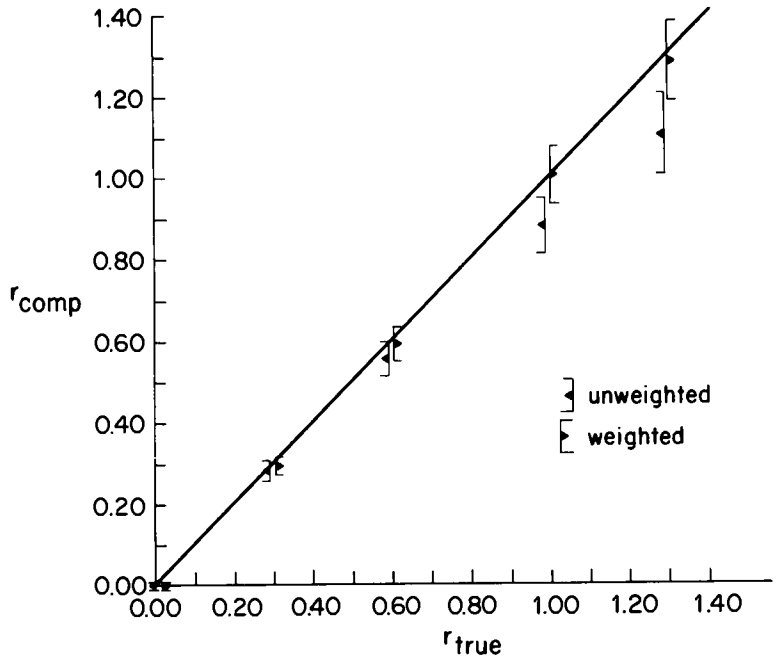


Fig. 4. r_{comp} as a function of r_{true}, for 5,000 expected counts per frame and range of liver motion of 20 rows. Means and 1 standard deviation ranges shown are averages for ten trials. For high values of r_{true}, r_{comp} is a biased estimator (unweighted values), but by using a modified algorithm which weights the data r_{comp} becomes an unbiased estimator.

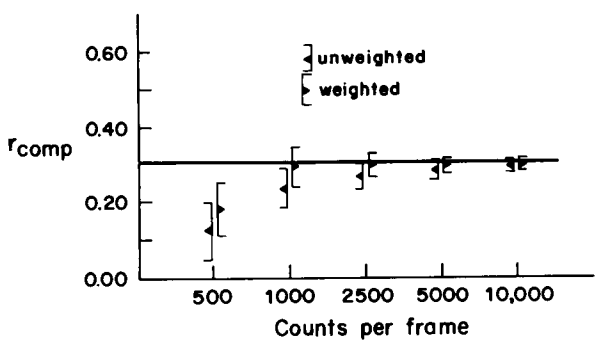


Fig. 5. r_{comp} as a function of expected counts per frame, for r_{true} of 0.30 and range of liver motion of 20 rows.

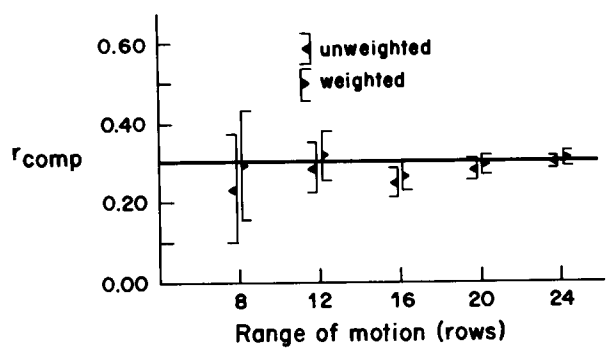


Fig. 6. r_{comp} as a function of range of liver motion, for r_{true} of 0.30 and 5,000 expected counts per frame.

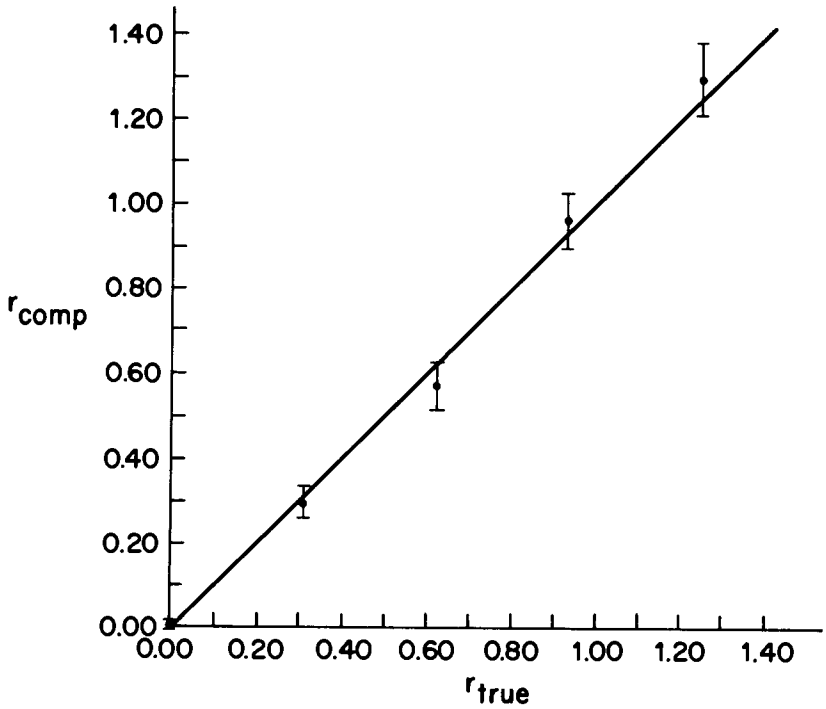


Fig. 7. r_{comp} vs. r_{true} for the moving liver phantom alone (r_{true} = 0) and with different combinations of background sources.

PHANTOM STUDIES

The stochastic method was tested by moving a liver phantom containing 2 mCi of ^{99m}Tc-pertechnetate beneath a Searle LFOV scintillation camera, and interposing various stationary sources between the phantom and the camera. The stationary sources consisted of a sheet source containing 0.4 mCi of pertechnetate, and six test tubes, each containing 0.2 mCi. Initially the true count rates from the phantom and each source were determined separately. Camera dead time effects, radionuclide decay and attenuation of one source by another were taken into account.

Five studies were carried out, consisting of the phantom alone, the phantom with the sheet source, and the phantom, sheet source and two, four and six test tubes above it. From the true count rates the values of r_{true} were found to be 0, 0.31, 0.62, 0.93 and 1.25 for these five studies. In each study the phantom was slowly moved back and forth over a distance of exactly 5 cm*, and a series of 100 one-second images was obtained. The stochastic method was applied to these images, using the "unweighted" curve-fitting algorithm. The resulting values for r_{comp} and its standard error are shown in Fig. 7. The agreement between r_{comp} and r_{true} was excellent (although had the "weighted" algorithm been used the value of r_{comp} would have been 0.06 higher for r_{true} = 0.93, and 0.10 higher for r_{true} = 1.25).

*This corresponds to 20 rows of a 128 x 128 array.

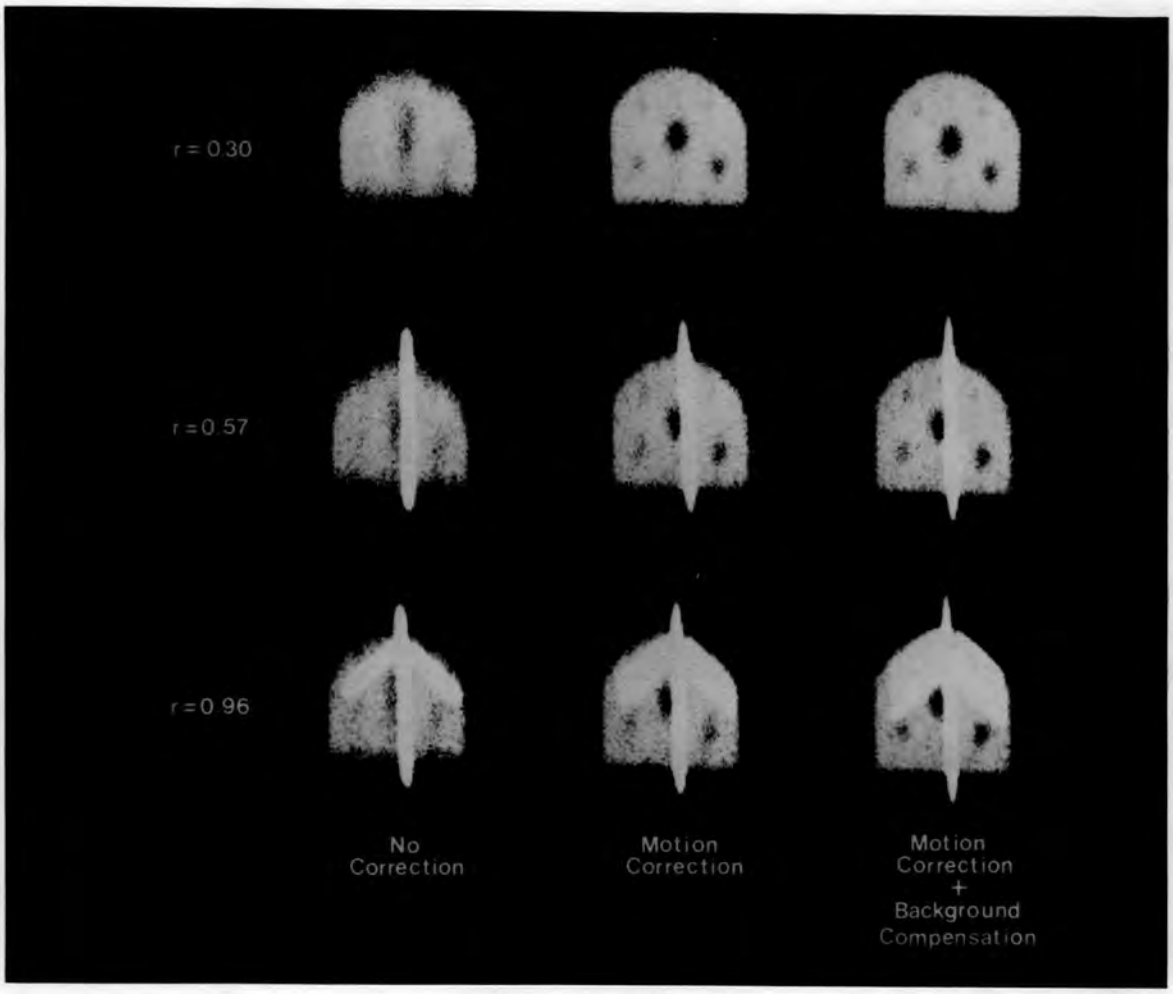


Fig. 8. Images from the phantom study showing the effect of adding background compensation to motion correction. A significant improvement is demonstrable only for $r_{comp} = 0.96$.

The images of the phantom were combined with and without motion correction by centroid fixation (Fig. 8). The values of r_{comp} were then used to add background compensation to motion correction (Fig. 8). If standard motion correction required that a frame be shifted a distance d in order to make centroids coincide, background compensation was added by increasing the shift to $d \cdot (1 + r_{comp})$. We see that for $r_{comp} = 0.30$ the improvement from background compensation is barely visible, for $r_{comp} = 0.57$ it is slight, and for $r_{comp} = 0.96$ it is definite but not striking.

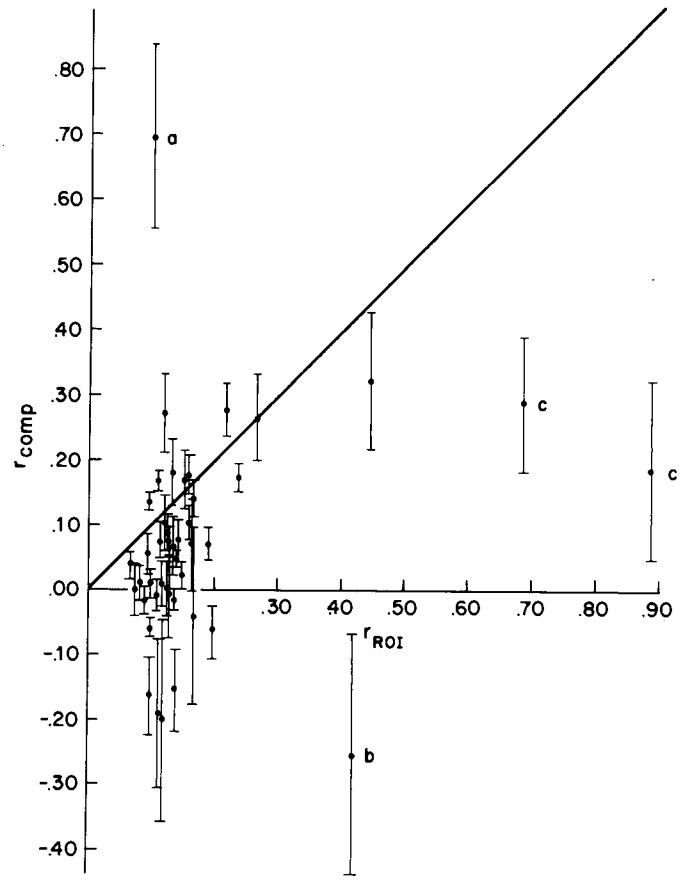


Fig. 9. r_{comp} vs. r_{ROI} for 44 patient studies. The value for the 45th study fell below the graph.

CLINICAL STUDIES

The stochastic method was applied in 45 studies of patients referred for liver scans at the University of Chicago. Following injection of 8 mCi of ^{99m}Tc-sulfur colloid a supine anterior view of the liver was obtained with a Searle LFOV scintillation camera. The patient was encouraged to breathe slowly and deeply, and 300-350 frames were collected at 0.3 seconds per frame. A count rate of the order of 3,000-6,000 per frame was usually obtained. The stochastic method was applied to the data to yield a value for r_{comp} with its standard error. The "region-of-interest" method was also used to estimate the background-to-liver ratio. This estimate is denoted by r_{ROI},

When r_{comp} was plotted as a function of r_{ROI} (Fig. 9) a rather poor correlation was observed. One value of r_{comp} fell too far below the graph to be plotted. The wide error bars associated with a number of values of r_{comp} were due in every instance to limited liver motion resulting from shallow respiration (see Fig. 6). This did not account, however, for the large discrepancy between r_{comp} and r_{ROI} in many instances, nor for the fact that some values of r_{comp} were less than 0, a physically impossible situation. We therefore reviewed our case material to try to identify the sources of error, concentrating particularly on the studies labelled a, b and c in Fig. 9.

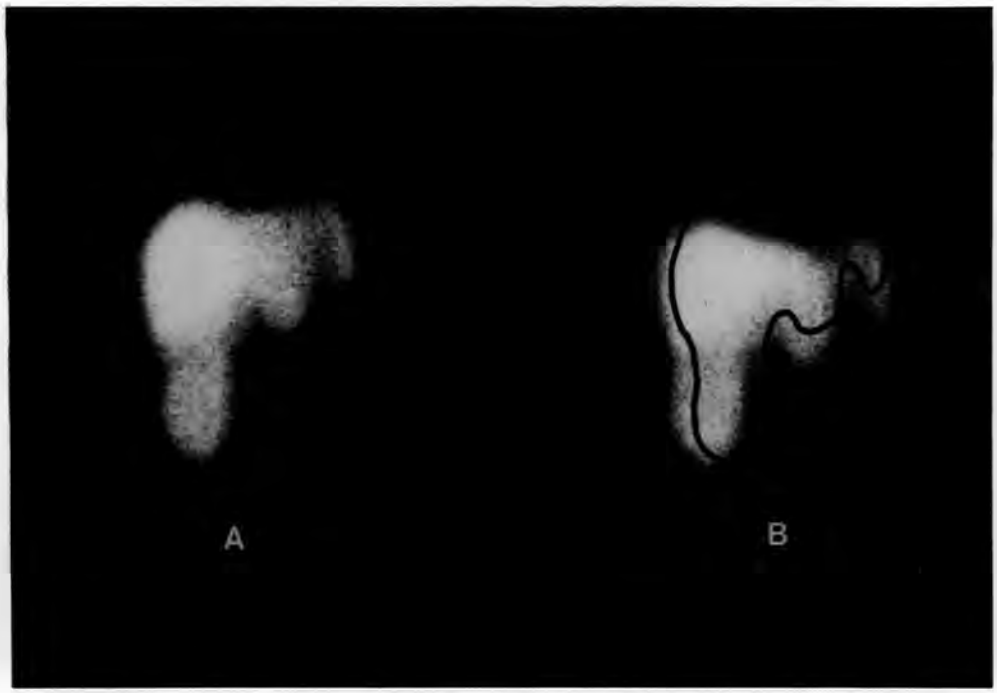


Fig. 10. Liver scan of patient denoted by "a" in Fig. 9. A. Expiration. B. Inspiration with superimposed outline of expiration image. The right lobe moves laterally in inspiration.

Study "a" was the only one in which r_{comp} was significantly higher than r_{ROI}. The patient had consolidation of the right lung with loss of motion of the right hemidiaphragm. While the spleen and left lobe of the liver moved downward on inspiration the right lobe moved mostly laterally (Fig. 10). Since the stochastic method does not recognize lateral motion it has treated most of the right lobe radioactivity as background, hence the erroneously high value of r_{comp}. Upon theoretical analysis we found that slight inhibition of motion of a portion of the liver does not significantly affect the value of r_{comp}, although the effect may become pronounced for marked inhibition.

The two studies denoted by "c" were the only ones in which a brain scan was performed just before the liver scan. The anterior view of the liver revealed radioactivity within the bowel, which contributed significantly to the background region-of-interest, leading to the high values of r_{ROI}. On the other hand, much of the activity could be seen to move with the liver, which tended to keep the values of r_{comp} low.

The other instances of significantly low values of r_{comp} are more difficult to explain. Case "b" involved an alcoholic with probable cirrhosis. The scan demonstrated prominent uptake in the bone marrow, hence, the rather high value for r_{ROI} was valid. It was noted in this and other cases with abnormally low values for r_{comp} that a small fraction of the radioactivity moved in and out of the field of view of the camera during respiration. By modelling this situation with the computer we discovered that a small amount of radioactivity moving in and out of the field of view can significantly depress the value of r_{comp}, even down to values below -1.00. We feel this is the probable explanation for the low values obtained in some of our clinical studies.

DISCUSSION

We have shown that the stochastic method for determining background-to-liver ratio gives reasonably accurate results in phantom studies. The results in clinical studies were less accurate, but we feel that greater care in centering the liver and spleen in the field of view of the camera and the utilization of a larger field of view would have circumvented the problem of moving radioactivity at the edge of the field of view, and this would have improved our results. We erred in presenting to our computer interface a field of view of only 32 cm, which is somewhat smaller than the maximum obtainable with the Searle LFOV camera.

The inspiration for this investigation was the desire to find an accurate and automatable method to compensate for background activity when applying motion correction in liver scanning. Fig. 7 demonstrates that such compensation is superfluous, since the improvement in image quality only becomes noticeable for values of r approaching 1.00. But of the 45 patients studied, only the two having brain scans prior to the liver scans had values of r (as measured by r_{ROI}) greater than 0.50, and these values were spuriously elevated.

This method might possibly be applied in the evaluation of hepatocellular disease, which leads to a shift of the radioactive colloid particles into the bone marrow. This shift is not very large,* however, and it is doubtful that this method would be as sensitive as other available methods.

We feel that the significance of this method lies in the fact that it illustrates the application of a statistical approach to an imaging problem. This approach has permitted the quantitation of the moving and stationary portions of an imaged radioactivity distribution, even though these portions are superimposed in the image and their separate configurations are unknown.

REFERENCES

1. B. E. Oppenheim, "A Method Using a Digital Computer for Reducing Respiratory Artifact on Liver Scans Made with a Camera," J. Nucl. Med. 12, 625 (1971).

2. P. B. Hoffer, B. E. Oppenheim, M. L. Sterling, and N. Yasillo, "A Simple Device for Reducing Motion Artifacts in Gamma Camera Images," Radiology 103, 199 (1972).

3. O. Fischer, "Method of Least Squares," p. 1295 in Survey of Applicable Mathematics, K. Rektorys, Editor, M. I. T. Press, Cambridge, Mass., 1969.

4. G. W. Snedecor and W. C. Cochran, Statistical Methods, 6th ed., p. 164, Iowa State University Press, Ames, Iowa, 1967.

*A much larger shift occurs to the spleen, but the method does not distinguish between liver and spleen radioactivity.

MODULARITY AND COMMAND LANGUAGES IN MEDICAL COMPUTING

A.E. Todd-Pokropek, D. Plummer and S.M. Pizer
University College Hospital, London ; University of North Carolina, Chapel Hill, North Carolina; C.E.A., Departement de Biologie, Service Hospitalier Frédéric Joliot, Orsay, France

ABSTRACT

In order to improve the user's interface with medical computer systems, the use of a command language is proposed. The definition of a suitable new command language, CL1, is given in detail. This language, which is not intended for processing images directly, but for controlling suitable processing program modules, has powerful features for handling conditionals, of great importance in such systems. The implications of the use of such a language on the architecture of such systems, and the constraints in implementing program modules within such a structure, are discussed.

To give an idea of how such a language could improve performance, results obtained with a subset in extensive use (here called CL0) are presented. Gains of up to a factor of 20 in throughput have been observed

INTRODUCTION

Medical computing, and medical image processing in particular, provides a particularly difficult environment in that procedures are not well defined and techniques highly likely to evolve. It has also become apparent that many of the established 'systems' used in this field are both highly limited and limiting. It is necessary to plan systems to adapt to this rapidly changing environment [1]. The aim of this paper is :-

a) to attempt to analyse the problems posed by the clinical usage of such systems,

b) to suggest an architecture which should improve the situation by being sufficiently flexible,

c) to make detailed proposals for a generalised implementation, and

d) to discuss results using one actual implementation.

This is a difficult topic to present, the 'problems' of using a system often only becoming apparent when some better user interface has been implemented. The first section on 'the problem' is included for completeness and should be skipped by the informed reader. The third section on the command language is the 'meat' of this paper, but is preceded by a brief discussion of architecture (Section 2) and followed by a discussion of the implications of the use of a command language on the form of the individual program modules (Section 4). Finally Section 5 presents results from a system using a subset of this language to indicate the benefits that should result, but it should be stressed that this implementation is very much a restricted form of the general command language proposed and should not be used as a model.

THE PROBLEM

To be clinically useful, a system should possess (at least) the following properties. It should be :-

a) simple to use

b) extremely flexible

c) error proof (resistant)

d) fast

and, of course,

e) it should perform useful tasks.

These requirements are often contradictory. If a system is flexible, this implies that the user has a great deal of control over individual operations. This generally means that the system is no longer simple to use. Probably the most fundamental problem is how to add power to a system, to increase flexibility and yet maintain a simple command structure.

It has been suggested that for use in a routine environment, a push button (invariant) type of system is best adapted, and that it is only in the research environment, where a physicist or computer scientist is available that 'programability' can be permitted. It appears to us that this distinction between such systems is false. The use of the system may be different, but the facilities that the system must offer are similar. The research application will benefit if the system is easy and fast to use, and the routine application can benefit if the predefined sequence of operations can conveniently be modified or extended.

In both cases, the way in which the system tends to be used is by executing 'protocols'. The protocol is a basic clinical entity, being that sequence of operations comprising one investigation or treatment. At first the sequence of operations required may be fairly straightforward, but inevitably, with time, it becomes longer and more complicated. A long sequence of manual operations becomes liable to error, <u>particularly</u> in the routine environment where the staff performing the study have less time per patient and are often less well trained than in the research environment. In most systems, far too much manual interaction is required.

Related is the problem of guiding the user through the system. Long descriptions accompanying each question tend to reduce the error rate but decrease throughput of the system. In addition, the situation frequently

occurs that the user is very familiar with one part of the system and therefore requires very little guidance, but is lost in another part and must be carefully routed through strange operations.

Unfortunately existing systems do not satisfy these needs to any great extent. 'Push button' systems, where the computer is carefully hidden, seem to be fast and after a certain period of training, 'simple to use'. They may be made relatively error-proof by interlocks. They are very inflexible. Although it is surprising how many clinicians are unable to type, most systems are controlled by keyboard input, often together with a joystick or light pen. However, whatever the 'hardware' used, the basic problem remains the same; defining a convenient user interface. The user is not very concerned with the internal form of a system, only how it reacts to him. Usually, this 'interface' is poorly designed, inconsistent, slow and clumsy (from the user point of view) and has no common convention (from the systems point of view) resulting in non-portable software. Most clinicians seem excited by the possibilities but frustrated by the facilities.

A more detailed analysis of the problem implies a proposed solution. The problem of 'ease of use' seems to arise because the user does not have a suitable language with which to address the machine. Error rate is related to the number of manual operations that the user is forced to make. The problem of non-portability seems to arise primarily from device specific code within a program module and imprecisely defined interfaces with the outside world. The problem of throughput seems to be related primarily to the time of 'response' of a user. Studies have shown [2] that in a typical medical application, the ratio of C.P.U. time to 'on-line' time is of the order of at least 10. Clearly this ratio will be smaller in certain applications where the processing is C.P.U. intensive, as in radiotherapy treatment planning. Concentrating on input (and to a lesser extent, output) transactions is a profitable way of speeding up such systems. The use of a form of 'batch' system, where all user interaction occurs before the bulk of the processing can improve performance, since the user is free to do other things and is therefore much less troubled by the time taken by any given operation. The problem of error correction and 'talking' the user through various operations is another manifestation of poverty in the input language available to the user.

ARCHITECTURE

The solution to this problem is to provide proper tools [3] for both the user and the programmer. This implies a certain structure or architecture for the system.

It is considered very important to isolate with well defined interfaces, the user, the operating system, the program modules and the peripheral drivers, as illustrated in Fig. 1.

The Operating System

Most machines function under an operating system. The operating system/user interface is the command string interpreter. To improve this user interface, functions such as loading device handlers, mounting disks, that is system specific operations, should be concealed as far as possible. A good technique is for the bootstrap to cause a file to be read containing all necessary initiation instructions. Similarly, the LOGON (or START) command at a terminal should perform all necessary system operations for

using that terminal. Originally, it was felt that clinical users should never be able to access the operating system directly. This was found to be a mistake. Use of the system is greatly improved if the clinical user can directly access the operating system command module in emergencies (when the system goes wrong) and for the programmer to have access whenever wishing to investigate the state of the system. This avoids duplicating a large number of facilities between the command language and the operating system.

A command language (except in the above circumstances) is a command string interpreter which will normally isolate users from the operating system but which will transmit requests from them in a standardised form to the specific operating system that is being used. Users therefore interact almost exclusively with the command language interpreter which acts as a shield between them and the rest of the system.

Utilities

Three types of program modules, other than the command language interpreter, comprise the 'system'. These are :-

1) the operating system

2) application program modules

3) utilities to connect program modules, command language and operating system.

In a well designed system, it should not be necessary to modify the operating system. The command language interpreter and the program module are both best written in a high level language. The key to good performance is the design of the utilities in which all system dependence should be incorporated. They should have standard interfaces to the application modules and command language. Section 4 of this paper is concerned with the design of these interfaces.

Program Modules

If the program modules are to be portable, they must perform precise, well-defined functions with all machine/system specific operations performed by 'utilities'. If standard data paths are used connecting modules to utilities and if the module is sufficiently general in the operation it performs, a powerful system will ensue. A library of useful modules can be created such that, by using various combinations, more or less any desired operation can be performed.

Giant all purpose modules should be avoided for many reasons. They are large, often slow, tend to have bugs and become 'idiosyncratic'. The system should be built of a number of small, very distinct application modules, connected together (in any reasonable manner) by the 'command language'.

However one factor which modifies this approach is the speed with which modules can be loaded. Three variations can be distinguished. When loading is slow, usually as a result of poor computer architecture (such that the loader is required to relocate addresses), larger modules performing several functions having a longer core residency seem inevitable to reduce the loss of time during loading. Better than the use of such large modules is the structure that tends to result when overlays are used. Effectively the linking between modules (or overlays) is created in the applications root program. Provided module loading time is small, this seems to be less

satisfactory than the proposed third solution in which these linkages are provided in the command language 'program', provided that module intercommunication is adequately supported as discussed below.

A fundamental implication of this third type of architecture is that control of the flow of modules must be provided in the command language. There is a temptation to construct command languages or protocol generators as macros where single commands are replaced by long text strings with parameter substitution. However, very rapidly, conditional branches, loops etc. are required. It is strongly recommended that the command language be constructed as a true language rather than grafting such facilities onto a form of macro command facility.

It must be emphasised that modularity, the separation of functions with well defined interfaces, reduces errors, interactions and improves performance. Thus the command language should be independent of the program modules, which should likewise be capable of running without using the command language facilities. The operator interacts directly with the command language which is an interface between him, the operating system and the program modules. The modules interface to the operating system (and the command language) by means of utilities. Thus the utilities are in the province of the system (machine level) programmer, and are machine specific, but with careful design and good utilities, the command language interpreter can be written in a high level language, perhaps even in FORTRAN.

Intertask Communication

Intertask communication has two aspects, those transactions needed for control, and those transactions involving data. The former can be handled by the command language, and will be discussed in Section 3. The latter have architectural implications. Usually, the quantities of data involved are substantial, and unless a good means of transmission between modules can be provided, modules will need to be concatenated, and module size will increase. At least three ways of transferring data can be implemented, and probably more than one technique will be required. These are basically: transfer by disk file, where data are written to backing store by the first module which informs a second module of the name of the file from which the data are to be recalled; transfer via core buffers, where data are placed in some common area of core for use by the second module; and 'pipelines'. A pipeline as developed in the "UNIX" operating system [3] is a channel using a small core buffer where data are fed from one module to another using a series of system provided flags to prevent data being used by the second module before having been created by the first module. Note that whatever the means used for 'bulk' transfer, modules do need to transmit small quantities of data (such as file names), a facility which must be taken into account by the command language.

In summary, the architecture of a desirable system as relevant to the subject of this paper is that implied by use of a number of program modules which are small, distinct, communicate via utilities and which are linked together by means of some command language.

THE COMMAND LANGUAGE

While the need for a command language is clear, what is less obvious is the scope of the language, and indeed, whether an existing general purpose language may suffice. It is suggested that a general purpose 'process control' type of command language is required, adapted for non-computer scientist

users but with very rudimentary arithmetic capabilities. It must be stressed that the proposed command language does NOT do any of the processing of data, but merely controls the flow of operations in the system. While FORTRAN is a reasonable choice for the application modules, (but see Section 4), it is totally inappropriate as a command language. Several other existing languages could be considered: PL1, CORAL, and in particular APL, but none of these seem to meet the requirement sufficiently. Thus while it is regrettable to devise yet another language for a specific application, little alternative seems to exist, unless other types of process control applications, being presumably faced with similar problems provide a suitable structure which could be implemented in the medical imaging area.

Basic Requirements

An ideal command language must not only have an acceptable syntax, but also be completely independent of the applications programs themselves. The aim here is to provide a command language which can be implemented above an existing set of modules with minimal modification to them and to the operating system. Perhaps a better feel for 'what a command language is' can be provided using a practical example. While a statement in a processing language might be of the form ARRAY(I,J)=MATRIX(I,J)/FACTOR, a typical statement in a command language could be RUN PROCESS,FILENAME=FRED. Users of big systems are familiar with such statements where they are generally termed part of the JOB CONTROL LANGUAGE. The command language is indeed closely related to job control, but in an interactive environment, designed for the non-specialist user.

The basic requirements in a command language are that:-

a) The syntax should be simple (and well documented) such that it is hard to make errors.

b) There should be continuity between crude and more sophisticated modes of use.

c) Error processing should be accessible and convenient (for example messages rather than codes should be issued).

d) The interface between the command language and the applications modules should be precisely defined and simple to use.

e) It should be easy to modify programs.

It is believed that a (ALGOL like) block structured language is more likely to satisfy these requirements. Jumps and GO TOs should be avoided, being inconsistent with a 'top down' design and more likely to generate side effects during modification [4]. A modular approach also greatly increases the ease with which simple functions can be expanded into more complex ones.

The main purpose of this paper is to propose and describe such a command language, based on previous experience. This imaginary language, following precedent, has been named Command Language number one, CL1, (or alternatively, Command Language for ON-line Execution, CLONE). A full version is currently in the process of being implemented at U.C.H. called, naturally, CLUCH. The existing subset has been renamed CLO.

Types of Transactions

Four types of transaction between the command system, the operating system and the applications program occur :-

a) unsolicited demand by the user, i.e. RUN a program

b) solicited requests of the user, i.e. question and reply

c) messages from the system requiring no action

d) unsolicited requests from the system (usually in the event of an error).

This last one only occurs when the program follows some unexpected course. It is important to take this possibility into account as the user may not be fully familiar with the program and thus be unable to predict its action.

The command language is designed to support all these facilities. The first (a) is implicit in the system 'RUN' being a command language instruction, (b) is achieved through 'ASK' which gets a value from the user and 'REPLY' which returns it to the program (as discussed below), (c) requires no direct command language action, although 'PRINT' may be used to generate local messages, and (d) may be processed either with an 'ON' which specifies the action to be taken on a particular occurence, or inherently by the system as any unexpected input request is referred to the user directly. Thus the command language can be placed completely between the user and the rest of the system, without restricting usage. In addition a small set of instructions is sufficient for performing basic operations. The new user does not have to learn an extensive repertoire of commands.

Blocks and Statements

The basic unit of a structured program is the block. Blocks here would be enclosed in begin, end symbols which may be BEGIN,END or more conveniently round brackets ( ). A block may be given a unique name by using a NAME: directive (see 3.9). Blocks may be nested up to any depth. A program must be at least one block deep, level zero being the console command level.

The following kinds of statements should be provided:-

1. assignment
2. conditional clause
3. repeat clause
4. on clause
5. transaction clause
6. command clause.

While the command language string itself should be pure, an associated impure 'work area' must exist in which values for use by program modules and transmitted from program modules can be stored. These should be stored by variable name defined in the command language. There is no direct connection between the name used by the command language, and the internal name used by the processing module with the important proviso that the command language name and the name specified by the module (when requesting input via the command language processor) must correspond as detailed in Section 4.1.

The Assignment Statement

An assignment statement is typically of the form:-

```
VALUE=100 .
```

This means that associated with the variable name VALUE is the string '100'. It was decided to store all data in string form (rather than as binary values) since all replies back to program modules can then be made in string (text) form, ignoring type conversions which can be performed by the program module itself.

A second very important form of the assignment statement, and probably the area in which the most problems in implementation exist, has the form:-

```
VALUE=VALUE+1 .
```

Here the string associated with the variable name value and the string '1' must both be converted into some binary value, operated upon (i.e. added) and then converted back into a string which is then reassociated with the variable name VALUE. The two areas of difficulty here are firstly the choice of an appropriate 'intermediate' form, here presumably integer, and the distinction between strings and variable names themselves. The following convention seems reasonable. Variable names never start with a number. In case of ambiguity, strings are surrounded by quotes. Thus :-

```
FILE='NAME',
```

means that the text string NAME is to be associated with the variable called FILE. The command language has never been thought of as a processing language, and the overheads imposed by the above constraints should not be serious. It should however provide normal arithmetic operations (+ - * /) with conventional algebraic operator binding and any depth of parenthesis permitted. Arithmetic operations on non-numeric 'items' are excluded, and replaced by concatenation of the text strings. For example:-

```
FILE='FRED'+'LIVER',
```

generates a text string FREDLIVER associated with FILE. More complex string handling would be performed by calling appropriate (external) program modules, and are therefore not defined immediately in the command language itself.

Conditional Clauses

It is very important to realise that conditional clauses are a vital part of a command language which otherwise would be of little value. The basic form of the conditional clause proposed is :-

```
IF logical expression THEN statement ELSE statement
```

for example:-

```
IF VALUE=0 THEN FILE='FRED' ELSE FILE='JANE' .
```

This has the meaning that if VALUE is equal to zero then the string FRED is associated with the variable FILE, and if not, the string JANE is

used. The boolean operators =,/=,<,>,<=,>= (or alternatively EQ NE LT GT LE GE), &, OR and NOT should be supported

To simplify the construction 'IF NOT logical expression THEN' an equivalent form :-

UNLESS logical expression THEN statement ELSE statement,

is proposed, for example:

UNLESS VALUE=0 THEN FILE='JANE' ELSE FILE='FRED',

is identical in operation to the statement given previously.

A well known problem with IF THEN ELSE is the ambiguity which arises if the ELSE is made optional (which is most desirable, to avoid the need for a null alternative statement). This can be overcome to an extent by the choice of a convention to determine the interpretation of nested IF statements.

Consider:

IF a THEN IF b THEN c ELSE d.

Two parsings are possible with statement d conditional on the negation of either a or b. The more common convention is to bind ELSE to the nearest 'un-elsed' IF, so making d dependent on b. However, the alternative of binding to the farthest IF (making d dependent on a) may appear more reasonable to the user unfamiliar with conventional programming languages.

The Repeat Clause

The repeat clause is another vital element of the command language. Its simplest form is :-

REPEAT statement,

whereby the statement is repeated indefinitely. This would be of little use however, unless extended by the use of conditionals. One important syntactical feature is that any statement can be replaced by a block. Thus the repeat clause can be re-expressed as :-

REPEAT (statement,statement.....).

Here the block defined by the begin and end symbols ( ) is to be repeated indefinitely. In the proposed language, four ways of controlling the repeat clause are suggested being the use of :-

1. WHILE
2. UNTIL
3. BREAK
4. LOOP .

The use of WHILE is in the form:-

WHILE logical expression REPEAT statement,

For example:

```
WHILE VALUE < 10 REPEAT (statement,statement, VALUE=VALUE+1),
```

which means that the associated statements are repeated if and only if the number associated with VALUE is less than 10. The logical expression should be evaluated before execution of the associated block. Thus if VALUE=10 the block is not executed, while if it is equal to 0 at the start, the block will be executed ten times.

UNTIL is very similar having the form:-

```
UNTIL logical expression REPEAT statement.
```

For example:-

```
UNTIL VALUE >=10 REPEAT (statement,statement, VALUE=VALUE+1),
```

which is identical to the WHILE statement given above.

Probably the most powerful manner of controlling a repeat clause is by use of the BREAK command, which is of the form:-

```
BREAK optional block name.
```

This means 'exit' from the block as named (that is execute the statement that immediately follows the named block) or, if unspecified, from the block in which the statement occurs. Thus the expression given above could also be written as :-

```
REPEAT IF VALUE>=10 THEN BREAK ELSE
   (statement,statement,VALUE=VALUE+1),
```

or alternatively :

```
REPEAT(NAME:XXXX,IF VALUE >=10 THEN BREAK XXXX ELSE
   (statement,statement,VALUE=VALUE+1))
```

where the command NAME: defines the name of the block as being XXXX. This facility enables BREAK to be used within nested blocks without ambiguity. Notice that if used in the lowest level block of a command language program, or when the name specified is that of the command language program itself, then BREAK is equivalent to RETURN as familiar in many programming languages.

Finally, the command LOOP is of the form:-

```
LOOP optional block name.
```

This has the affect of forcing the program to the end of the named block, which, if the block is under the control of a REPEAT with a WHILE or UNTIL will cause re-evaluation of the logical expression. Thus the expression given above can also be written as:-

```
WHILE VALUE <10 REPEAT(statement,statement,VALUE=VALUE+1,LOOP),
```

where the LOOP is essentially dummy, or else:-

```
REPEAT(IF VALUE<10 THEN (statement,statement,VALUE=VALUE+1,LOOP)
   ELSE BREAK).
```

Normally the LOOP command will depend on some condition and a more useful example is of the form:-

```
REPEAT(NAME:XXXX IF VALUE<10 THEN(statement,statement,VALUE=VALUE+1,
   IF NUMBER=0 THEN LOOP XXXX ELSE statement,statement)).
```

Note that:

```
REPEAT(statement,statement,VALUE=VALUE+1,IF VALUE>=10 THEN BREAK) ,
```

is not identical to the other forms given above since, if on entry VALUE is greater or equal to 10, the statements are still executed which was not true for the other expressions.

The On Clause

The ON clause proposed may be familiar from PL1. It provides for the processing of a 'software interrupt'. Thus the occurrence of a particular event can be detected and any necessary action taken asychronously with the main program. This facility is particularly useful in error processing, as a single statement can have the effect of a test for error after each operation.

The general form of the statement is:-

```
ON logical expression THEN statement.
```

For example:-

```
ON ERROR=3 THEN STOP,
```

where STOP is a command to halt execution of a command language program. This ON clause is interpreted as meaning that if at any time the value of the variable ERROR becomes equal to 3, the program will cease execution.

The 'statement' may of course be replaced by a block and more complex operations performed which will frequently be necessary. In particular, two facilities are needed, to permit return to the interrupted portion of the command language program, and to permit a return to some other (usually earlier) point in the command language program. It is assumed that if the statement following the THEN is completed, control is returned at the interrupted part of the command language program. The second facility could be provided by a JUMP statement, but this seems to be unnecessarily 'general', makes the operation of the command language processor rather more complicated, and introduces considerable scope for making errors. Instead it is suggested that the command RESTART of the form:-

```
RESTART optional block name
```

be used, where obeying this command causes execution of the first statement in the named block (or if no name is specified, the block in which the command is issued, excluding any ON blocks).

Thus an example of an ON using this facility could be:-

```
ON ERROR=3 THEN (statement,statement, IF ATTEMPT >3
   THEN STOP ELSE RESTART).
```

The ON statement is of most use in error handling, a complex subject which will be considered in a little more detail later.

Transaction Clauses

Transaction clauses affect the transfer of text string between the user, the command language processor, and the program modules. They should include the following types of statement:-

1. ASK
2. REQUEST
3. REPLY
4. PRINT
5. GET .

The ASK statement followed by an optional prompt string has the following effect. When a program module wishes to input a value (ask a question), if the next statement to be executed in the command language program is an ASK, then the user will be directly asked the question, and the command language processor will pass to the next statement following the ASK command (see also Section 3.10). The prompt string if given overrides the prompt in the program module.

It is suggested that ASK could also be used in another rather different manner by being included in an assignment statement in the form:-

```
Variable=ASK optional prompt string.
```

For example:-

```
FILE=ASK ,
```

or:-

```
FILE=ASK 'WHAT IS THE FILE NAME' .
```

Thus if an input request for the named variable occurs, then the input request will be passed to the user with the prompt supplied by the program module in the former case, and the prompt supplied by the command language in the latter. The prompt string could be a variable name (in which the prompt has been placed) and therefore in this example, the string quotes are required.

The command REQUEST replaces ASK in the above case when the request to the user is to be made immediately, as opposed to when and if the variable is requested by the program module, for example using:-

```
AGAIN=REQUEST 'DO YOU WANT TO CONTINUE' ,
```

the question would be asked immediately, and the reply placed in the variable AGAIN. This might well be followed by the conditional statement:

```
IF AGAIN='NO' THEN BREAK ELSE LOOP.
```

Thus a means is provided for the command language processor to ask questions of the user itself, and to take action dependent on the reply.

The command REPLY has the form:

```
REPLY string or variable name,
```

for example:-

```
REPLY 100   (or '100'),
```

or alternatively:-

```
VALUE=100
REPLY VALUE.
```

This command has the following effect. If the next statement to be executed in the command language is a REPLY, the next input request made by the program module will have the associated string returned as its value, and the command language processor will commence execution with the statement following the REPLY.

Thus suppose that a program makes two input requests, the first for FILE being a file name, and the second for FRAME being a frame number. Predefined values could be supplied by either:-

```
FILE='FRED'
FRAME=2 ,
```

where the order is not important, or:-

```
REPLY 'FRED'
REPLY 2 (or '2') ,
```

where the order is important, but the variable names themselves are not required.

There is an associated distinction between these two forms of specifying values related to synchronisation which will be considered in Section 3.11.

The order of looking up for a value is very important. There is some scope for modifying the action dynamically firstly when the command system is not active but also to control the return of default values specified by the module when there is no user defined reply. The order should be:-

If the command language system is not active, steps 1, 2 and 3 are skipped.

1. Check the list of variables in the impure area. If the name is defined, use the associated value. Note that the value may be ASK in which case a direct request to the user is generated.

2. If a default value has been provided by the module for the request, this value is returned. A SET option (Section 3.9) could be provided to modify the conditions for the return of default values.

3. If the next statement in the command language program is an ASK or REPLY, perform the appropriate transaction.

4. When all else fails, pass the request to the user.

The command PRINT has the form:-

PRINT text string .

For example:-

PRINT 'NOW TURN THE PATIENT OVER',

which has the effect of printing the associated message on the users console.

The command GET has the form:-

GET variable name.

For example:-

GET COUNTS,

which has the effect of saving for future use the value associated with the variable name AS SUPPLIED BY THE PROGRAM MODULE. This implies that at some point in the execution of the program module, a call is made to the command language processor (as specified in Section 4.4) of the form:-

CALL RECORD('NAME',value).

The value is thus associated with the variable name, and may optionally be saved by the command language program by using the GET command. This provides a means for the program modules to communicate data to the command language program. All module calls of the form RECORD are placed in a pool for use by the command language. On a GET specifying a variable name not so far in the pool, the command language suspends until the value is specified or until the program terminates, when the pool is emptied.

The Command Clause

The command clause is basically a statement instructing the command language processor to take some action and includes:-

1. CALL
2. RUN
3. PAUSE
4. RESUME
5. STOP
6. BREAK
7. LOOP
8. RESTART
9. NAME:
10. SET

plus a whole series of statements which only have meaning at 'console' level

(that is - not within a command language program) being:-

1. EDIT
2. LIST
3. CREATE
4. DELETE
5. STATUS
6. etc. etc.

The form of the CALL command is:-

CALL command language program name,

and has the effect of causing execution of the command language program as named. The word CALL is optional. Thus:-

CALL PROCESS,

and

PROCESS,

are equivalent. The name of a command language program is the name of the outmost block.

The form of the RUN command is:-

RUN program module name,

and has the effect of causing execution of the named program, and starting a block in the command language program of the same name (unless explicitly renamed with NAME:). As above, the word RUN is optional. Thus:-

RUN COLLECT,

and

COLLECT,

are equivalent, the order of looking up a command being:-

1. Is the next word a reserved word, that is, some word having meaning within the command language such as RUN, IF etc.

2. Is this the name of a command language program

3. Is this the name of a program module.

Thus a typical use of a RUN command could be:-

COLLECT(COUNTS=1000,TIME=100).

PAUSE is a command suspending execution of the currently running command language program. It may be issued either within the program, or as a terminal command. The impure area is left untouched. The command language program may be resumed with the command RESUME. In between the two commands,

the system may be used for any other operation, which is particularly useful for handling errors, and debugging.

The commands STOP BREAK LOOP RESTART have the meaning detailed previously. However they may also be used as console commands after PAUSE with appropriate meaning.

The NAME: command has the form:-

NAME: name,

and may be used for naming blocks and command language programs.

The SET command has the effect of specifying values for certain system variables and has the form:-

SET options statement,

For example:-

SET HELP=1,

to control the level of prompting.

This is included to enable the command language to cope with some system specific functions and has been left deliberately vague.

In addition to the above commands, there is clearly a set of commands which have little or no meaning within a stored command language program but must be available at console level. The operation of the system at console level is slightly different from stored program level as is discussed below. The commands referred to are primarily needed to manipulate command language programs themselves and would include such functions as EDIT LIST CREATE DELETE, all followed by an optional command language program name.

The command STATUS followed by an optional statement is the equivalent operation to SET, permitting system specific functions to be output.

Console Level Commands

The operation of the system at console level is important and in implementation not necessarily simple. Two rather opposed facilities are required. Firstly, the commands and statements having meaning within a command language program should also have meaning when typed at the console. For example:-

COLLECT(COUNTS=1000,TIME=100),

should, as in the stored program example above, force the program COLLECT to run, with predefined values for the two parameters. It is suggested that a good implementation would be to store this 'command' as a normal stored command language program (probably named CURRENT or equivalent) and then execute it, such that the command language program so generated would be available and could be stored permanently afterwards if required. The first major difference suggested between console level programs, and stored programs is that it is considered desirable to treat the 'carriage return' at the end of the command line as terminating the block and forcing execution (unless the user indicates by leaving a trailing comma that more instructions are to follow).

Secondly, for convenience, several special command language statements may be required, for example:-

DO 2 COLLECT(COUNTS=1000,TIME=100),

meaning DO, or repeat, the block that follows 2 times. This form of a repeat clause is not really required in the stored language since it can be provided in many other more elegant forms, but it is a very compact and simple form of a commonly required operation.

The second (rather contradictory) requirement is that the system should be usable and function normally without the command language processor. When a program requests input, the sequence of operations, as given in Section 3.8, should be obeyed. Note that this system functions correctly in the absence of the command language processor, which could then be 'grafted' into a pre-existing system.

Additionally, it is desirable if the console level commands for the command language processor resemble the normal commands for the operating system. Thus the command:-

RUN PROGRAM,

will behave correctly in many operating systems, and be interpreted correctly by the command language processor. It may be necessary to provide some 'aliases' for example R as equivalent to RUN etc. to improve the visual appearance. Transparency of the command language processor can be improved (at console level) by simply passing to the operating system unrecognisable strings, such that all special system specific operations would still function normally.

Synchronisation of Command Language and Modules

The need to synchronise operations in the command language processor and program modules is very important and can be achieved by obeying a few simple rules. Firstly, a RUN command is only executed when the module previously requested has terminated. This can be achieved either by the module sending a message to the command language processor stating that it is about to terminate, or for the command language processor to look at the system at regular intervals to see if the program module is still running. The former system is slightly preferable, since the modification of the module required can serve other functions such as closing files etc. Secondly, the statements ASK REPLY GET all cause the command language processor to suspend. When an input request is received, the next ASK or REPLY is executed (but see Section 3.8) and as before the command language program continues to run until the next synchronising statement (ASK,REPLY,GET,RUN). A GET command is similarily synchronised by a RECORD request from the program module. For generality, it is considered that a special SYNCH command be added to the language having the form:-

SYNCH optional name ,

with the following property. SYNCH is a synchronising statement and causes the command language processor to suspend when encountered. When no name is specified, the next input request or RECORD request causes the statement to be skipped. When a name is specified, this implies that this statement is skipped only when the program module makes an input request or a special CALL SYNCH('NAME') request specifying that name.

The operation of this system is best illustrated by example. Suppose that a contour program requests a series of contour values. A portion of the command language program for handling this could read:-

```
RUN CONTOUR, LEVEL=0,WHILE LEVEL<1000 REPEAT LEVEL=LEVEL+100.
```

When the previous program had completed, the CONTOUR program would be requested, and the repeat clause executed until LEVEL takes the value 1000. Thus, the first input request from the program for LEVEL would receive the reply 1000. This is clearly undesirable and may be corrected by writing:-

```
RUN CONTOUR, LEVEL=0,WHILE LEVEL<1000 REPEAT(LEVEL=LEVEL+100,REPLY LEVEL),
```

or alternatively:-

```
RUN CONTOUR,LEVEL=0,WHILE LEVEL<1000 REPEAT(LEVEL=LEVEL+100,SYNCH).
```

In both cases, the first value returned is 100, ten values in total being generated in steps of 100.

Separators

In any real language, separators have to be carefully defined. Here the following rules are suggested. Spaces and 'terminators' complete 'words'. Terminators include: new line, comma, semi-colon, open and close parentheses. New-line has a special meaning at console level. Logical expressions are terminated by the reserved word THEN. Strings are surrounded by string quotes (or are numeric), but strings are also terminated by new-line. Comments follow a semi-colon and are terminated by a new-line, or a second semi-colon. A special 'concatenate string symbol', double quotes, needs to be provided to enable strings to extend beyond one line. Arithmetic operators are handled conventionally. Statements are either terminated by a comma, semi-colon or parenthesis, or, alternatively by a new-line, except when the statement is syntactically incomplete and can be extended by use of the following line. All examples given in this text follow these rules. The use of parentheses in expressions is permitted and follows the same rules, blocks returning values being created. A proposed syntactical definition is given in Appendix 1.

Implementation

While it is not anticipated that the complete language will be implemented entirely as described here, some guide-lines are necessary. It is considered highly desirable that the command language processor be an interpreter. This means that the command language program is stored in a form directly representing the original commands. It cannot be compiled into machine code as such, since it consists primarily of data such as program names, etc. There is little advantage in compilation in any case, since it is NOT a processing language, but used for controlling modules, and there would be very little gain in speed. Compilation would entail the existence of source and 'execution' type modules, which only complicates editing and housekeeping. One of the main benefits of a command language is the ease of modification and avoidance of any compilation phase, so that changes can be made at 'run time'.

Storing the program directly as text seems undesirable, and considerable compression without loss of the ability to reconstruct the original input can be achieved. Avoiding jumps and making all internal variable references by name removes any need for complex processing. In implementing the run-time

system, it is necessary to take care to compact the impure area, and to perform efficient garbage collection of no longer required variables. However, the command language program can be read sequentially, instruction by instruction, and only a very small buffer is required for the interpreter. No problems have been experienced in implementing the CLO system in a multi-tasking multi-user environment.

THE PROGRAM MODULES

Within the structure discussed previously, program modules should have certain properties. They should:-

a) be small

b) have a well defined interface

c) be 'structured'

d) use utilities for user and system communication.

The structures so far outlined were designed to coordinate modules with these characteristics. It is necessary however to define the communication utilities. These are clearly system dependent and should contain all that code which is device and system specific and thus renders modules non-transportable.

Structured programming [4] has already been discussed and is clearly desirable. After some hesitation it was considered that despite the drawbacks associated with FORTRAN (being unstructured and poorly adapted to present machine architecture), it still remains the language of choice. It is quite reasonable to treat FORTRAN as a machine independent intermediate code to solve the portability problem. A super set of FORTRAN is then required which can be pre-processed into FORTRAN. One example, for which a FORTRAN pre-processor is available, is RATFOR [3]. RATFOR permits block structures, conditionals, reasonable loops and eliminates the unpleasant FORTRAN line based syntax. As presented, it also avoids FORTRAN READ and WRITE statements. The remaining description of the requirement for program modules will be expressed in FORTRAN terms noting that it could be adapted relatively easily for some other environment.

Input Transactions

Input output transactions are principal areas of interaction between program modules and the command language. Clearly the READ statement must be eliminated, not only because of restrictive formats but also as few systems will allow FORTRAN input to be intercepted by a command system. The proposed form for input requests is:-

1) I=VALINT('variable name', 'prompt'.....)
   to input an integer value I

2) R=VALREA('variable name','prompt'.....)
   to input a real value R

3) L=VALSTR('variable name','prompt',array,max.length.....)
   to input a string to an array where L receives the length of the string placed in 'array'

4) I=VALDIR('variable name','prompt',directory.....)

An additional facility which has been found very useful is to look-up a string in a directory and return an associated integer value. This can be implemented using the above call.

Optional parameters (indicated above by '.....') are important. As actually implemented, three optional parameters are provided for each call, being a default value, a lower limit, an upper limit. The provision of these facilities not only reduces module size and complexity but encourages the programmer to include checks which might otherwise have been omitted.

The format of these requests depends heavily on the ability of the system to handle a variable number of parameters in a call. It is desirable for a novice user to be able to make a simple call using minimum parameters such as I=VALINT. The more skilled user can take advantage of the more general form. Unfortunately, many FORTRAN implementations provide no means of extracting the number of actual parameters passed in a call and in some, variable length lists actually cause errors.

Output Transactions

The justification for eliminating the FORTRAN WRITE statement is less obvious. Devices used in medical systems are often inconvenient to use from FORTRAN. For example, when writing text to a colour television, it is difficult to specify in FORTRAN the position and colour in which the text is to be displayed. It is felt that the use of sub-routines for performing output is preferable. The following calls are suggested:-

1) CALL OUTINT (value, no of characters.....)
to output an integer

2) CALL OUTREA (value, no of characters, number of places after decimal point...)
to output a real

3) CALL OUTSTR (array, no of characters)
to output text

4) CALL TAB (column)
to tabulate to a particular 'column'.

All the above put text into a buffer. An extra call is needed to output the buffer, CALL NEWLIN. The 'number of characters' parameter should be optional. If unspecified, the system should use suitable default values, type conversion being performed automatically. Carriage control (line positioning) etc. can be added as a further optional parameter in the NEWLIN call. Note that this system is in fact much more flexible than FORTRAN WRITE statements in that back-spacing, character substitution etc. are easily performed.

File Handling

An area of vital concern is that of file handling. Unfortunately file handling is VERY system specific. However, by suitable use of subroutine calls, and the use of a system specific data array (here called the file descriptor in which all the pointers, buffers, device names, etc. are stored), reasonable portability may be obtained. It is considered that all I/O should be treated as coming from or going to a file structured device. In order to

set up the file descriptor, a number of subroutines will be necessary such as:-

```
CALL DEFFIL(file descriptor,......) ,
```

to define the file, with a number of optional arguments, and

```
CALL FILNAM(file descriptor,'name') ,
```

to define the name of the file.

Secondly, it will be necessary to open the file with a call of the form:-

```
CALL OPEN(file descriptor).
```

Since more than one file at a time may be opened, it is desirable to add a call to select a particular file as being the 'currently active' file by:-

```
CALL SELECT(file descriptor) .
```

A further call to close the file (although any reasonable system should close all open files when the program terminates or aborts) is:-

```
CALL CLOSE(file descriptor) .
```

A more detailed discussion of file handling is beyond the scope of this paper, but from the point of view of most medical imaging operations, the file handling as required for matrices should be concealed within some utility, for example:-

```
CALL MATRIX(array, length, file descriptor,.....) ,
```

to input and output arrays of numeric data.

Command Language Communication

As described in Sections 3.8, 3.11 (GET,SYNCH), some communication between module and command language is necessary. RECORD places a variable name and value in a command language processor pool from which it can be extracted by the GET statement, specifically:-

```
CALL RECINT('name',value) to be used for an integer
CALL RECREA('name',value) to be used for a real
CALL RECSTR('name',value) to be used for a string .
```

Programmers should consider which variables may be useful to the command language and provide appropriate RECORD calls for them. The SYNCH call is:-

```
CALL SYNCH('name') ,
```

which will cause the SYNCH statement with the given name to be skipped in the command language program. When the command language processor is absent, these calls have no effect.

Two extra calls are both necessary and convenient being:-

```
CALL START (optional parameters)
CALL FINISH (optional parameters).
```

These routines should be called at the beginning and end of every program module and can be used to perform a variety of system dependent tasks such as attaching the correct input stream from the user to the module etc.

Help

Using the command language processor, and the module calls as given here, it is relatively easy (for the command language processor) to print out the name of all variables in the program modules as reference is made, simplifying the writing of command language programs. A second much more important feature enabled is that of good general purpose 'help' facility. It is considered that various levels of 'help' should be provided. The level under which the system operates would be dependent on the experience of the user. Thus the lowest level of 'help' might be that, when input is required, the variable name is prompted. The next level of 'help' would use the full prompt as included in the call for input. A third level of prompting could be provided by using the command language processor to associate with an input request a length of text stored in some associated file. This would be created by the programmer (or user) running the program in some special mode such that for each input request, the corresponding text could be typed and saved in a file. It seems highly advantageous to perform this at run-time rather than while programming since the much greater convenience that this offers suggests that programmers might actually make use of this facility.

For example, suppose that a program made a call of the form:-

```
A=VALREA('PARM1','INPUT THE COEFFICIENT OF FATTINESS'),
```

with associated text:

```
THE PROGRAM NOW REQUIRES THE COEFFICIENT OF FATTINESS
THIS IS A REAL NUMBER EQUAL TO THE PATIENTS HEIGHT
DIVIDED BY THE PATIENTS SURFACE AREA AND CAN BE FOUND
DESCRIBED AND TABULATED IN DOCUMENTA GEIGY ON PAGE 143
NOW INPUT THE VALUE.
```

Thus help level 1 prints:-

```
PARM1?
```

Help level 2 prints:-

```
INPUT THE COEFFICIENT OF FATTINESS,
```

and level 3 prints the full associated text.

Further efficiency can be achieved by suppressing prompts in replies when typed ahead, enabling or disabling the use of default values, and adding suitable clarifying messages when errors are detected.

Error handling, a tricky subject, can be improved by the use of such calls in a command language environment, since indications of acceptable

ranges and default values can be printed, and indeed, the help level can be altered dynamically.

RESULTS

The proposed command language is based on and modified from an implementation written on the Informatek Simis 3 computer system [5], here called CLO, which was based on an implementation of a pure macro processor (replacing long sequences of commands by a single command for a Varian 620i computer). Various control functions were grafted onto the Informatek system such that in terms of capability it has most of the facilities described in the command language CL1 above, but it is much more cumbersome and is full of gargoyles (software bugs described as 'features' rigidly built into the system). It is not even possible to remove these gargoyles by improving the structure of the command language since enormous numbers of user defined programs would have to be recoded. The input requests in the modules did not include any variable name and therefore assignment presented certain difficulties. The trick used was to treat the first four characters of the prompt message as the variable name, which eliminated almost all ambiguity.

A second major defect is that this implementation is more of an assembly language (IBM JCL is another example) than a high level language. However, the simple user needs to learn but five commands (RUN,REPLY,ASK,PRINT,STOP) and can immediately write valid, simple command language programs. As users become more experienced, flow control, conditionals etc. etc. become necessary and incorporated in their programs. The process of learning to write command language programs can be quite gradual. This is extremely important since clinical users rarely have the time, desire, or patience to attend training sessions or to read manuals except when the need is urgent. In fact, clinicians, physicists and computer operators have written CLO command language programs.

Several interesting observations can be made about the use of a command language, which are not intuitively obvious. Firstly, programs tend to be long. Most typical CLO command language programs consist of the order of 300 statements, although some are very much longer than this. Secondly, about half of the statements are concerned with control of flow and half with running programs and defining replies. This illustrates that clinical protocols tend to become quite complicated with many conditional paths and loops. Thirdly, the running time is often substantial. A typical CLO command language program takes at least 15 minutes to complete. This suggests that pre-defining parameters or running a batch type system for processing overnight can be extremely useful.

Two areas where considerable effort was unexpectedly required were editing and debugging. A reasonable editor is essential and should be made simple enough for a clinical user. Debugging is even more important. It has been remarked that in many large systems users require more time to debug the job control language than the program itself [6]. The command language CLO as implemented originally proved very troublesome in that error correcting, modifying and checking were very time-consuming. This is primarily a result of the long running time of command language programs. The solution adopted was to add line numbers to the command language string which could be used together with a trace facility. In trace mode the command language program could be stopped and re-started externally at any desired statement number, variables output and modified, the interpreter made to function 'step by step', printing out each operation, etc. A combination of these facilities considerably reduced debugging time. The single most important facility was surprisingly

the ability to jump to a particular point in the program, skipping all the previous (tested) operations, only executing that part of the program in which the error occurred. The errors observed were almost exclusively errors in the predefined values and logical errors in the command language program resulting in a different sequence of operations to that desired by the user.

A comparison was made between operations performed manually and using the command language. The results were dramatic. The error rate was extremely low when using the command language. The time of execution improved enormously (depending on the nature of the task) from a factor of 3 at worst to a factor of 20 at best. This is a highly significant improvement. It should also be stressed that with a command language program, the system is considerably easier to use. Contrary to anticipations, CLO command language programs are more commonly used in centres dominated by routine than those with a more research orientation, probably because command language programs significantly increase throughput.

A further unanticipated benefit, presenting very interesting potential, is that command language programs now represent clinical protocols in a formalised manner. Transfer of protocols between centres is therefore feasible even though this may require the translating or pre-processing of command language programs. Recently, there has been more discussion about exchange of CLO command language programs than of application modules.

A similar system to that existing on the Informatek system has been implemented on a PDP11, and is in the process of being extended to execute the full CL1 command language as described here. Both systems can operate independently of the command system and follow the basic structure shown in Fig. 1.

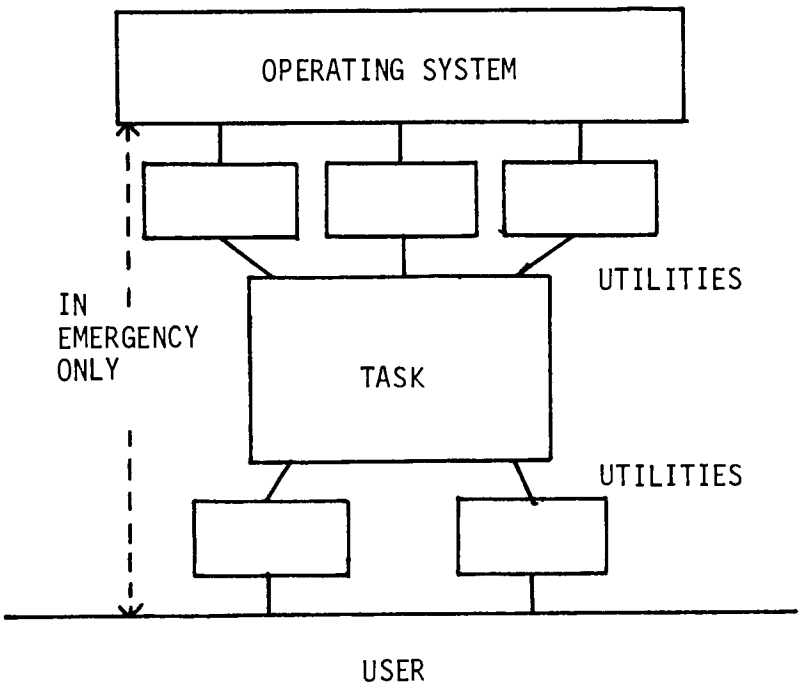


FIG. 1. The separation of the user, the operating system and the program modules (tasks) by utilities.

As an illustration of practical command language programs, Fig. 2 shows an example of CLO, being one subroutine in a renogram processing command program, as typed by a user on an Informatek system. LOAD is a synonym for RUN; * indicates a comment, here used as a name; TELZ MASK ADDY and COPY are four program modules. It makes calls to other command language subroutines called 57,44,50,58 and 42. Fig. 3 shows the same program as listed on the Informatek system. QUESTION is synonymous with REPLY, the IF statement asks a question (REQUEST) and takes one of three branches dependent on whether the reply made was 'Y', 'N' or 'T'. Assignment has not been used in this example being a recent upgrade in CLO. Fig. 4 shows the same program translated literally into the command language CL1 described here. The names of the subroutines called have been prefixed by the letters MAC for clarity. Finally, Fig. 5 shows the same program using assignment, making use of implicit replies and default values.

```
CREATE

*,MAC 12 SUB 10

IF,DO YOU WISH TO MASK THE BLADDER?

THEN,ELSE,ELSE

(,LOAD,TELZ(104,Y,,1,,4,)

LOAD,MASK(1,27,104,1)

LOAD,ADDY(1,27,99,,)

LOAD COPY(99,1,105,),),CALL,57

IF,TO FIND KIDNEY ROI'S, TYPE "Y" FOR ISOCONTOURS; "N" FOR

BUG, METHOD

44,50,50,CALL,58

IF,DO YOU NEED TO TRY AGAIN TO DETERMINE THE KIDNEY ROI'S?

42,CONTINUE,CONTINUE,RETURN,STOP
```

FIG. 2. Input for command language subroutine in CLO.

```
MACRO NO.   42
    1--, *COMMENT=, MAC 12 SUB 10 .
    2--, IF=, DO YOU WISH TO MASK THE BLADDER? .
        FOR Y OR +, CONTINUE THEN,
        FOR N OR =, ELSE,
        FOR T OR -, ELSE.
    3--, ( OPEN PARA .
    4-----, LOAD=, TELZ*****************************************************,
    5-----, QUESTION=, 104 .
    6-----, QUESTION=, Y .
    7-----, QUESTION=, .
    8-----, QUESTION=, 1 .
    9-----, QUESTION=, .
   10-----, QUESTION=, 4 .
   11-----, QUESTION=, .
   12-----, LOAD=, MASK*****************************************************
   13-----, QUESTION=, 1 .
   14-----, QUESTION=, 27 .
   15-----, QUESTION=, 104 .
   16-----, QUESTION=, 1 .
   17-----, LOAD=, ADDY*****************************************************,
   18-----, QUESTION=, 1 .
   19-----, QUESTION=, 27 .
   20-----, QUESTION=, 99 .
   21-----, QUESTION=, .
   22-----, QUESTION=, .
   23-----, LOAD=, COPY*****************************************************,
   24-----, QUESTION=, 99 .
   25-----, QUESTION=, 1 .
   26-----, QUESTION=, 105 .
   27-----, QUESTION=, .
   28-----, ) CLOSE PARA.
   29--, CALL=, 57.
   30--, IF=, TO FIND KIDNEY ROI'S, TYPE "Y" FOR ISOCONTOUR; "N" FOR BUG, METHOD
        . ,
        FOR Y OR +, CALL, 44,
        FOR N OR =, CALL, 50,
        FOR T OR -, CALL, 50.
   31--, CALL=, 58.
   32--, IF=, DO YOU NEED TO TRY AGAIN TO DETERMINE THE KIDNEY ROI'S? .
        FOR Y OR +, CALL, 42,
        FOR N OR =, CONTINUE THEN,
        FOR T OR -, CONTINUE THEN.
   33--, RETURN=.
   34-, STOP=.
EOM
```

FIG. 3. Output from the command language program listing as defined in Figure 2.

```
NAME: MAC12  ;subroutine 10
IF REQUEST 'DO YOU WISH TO MASK THE BLADDER'='Y' THEN
     (RUN TELZ, REPLY 104, REPLY 'Y',REPLY,REPLY 1,REPLY,REPLY 4,REPLY
      RUN MASK, REPLY 1,REPLY 27,REPLY 104,REPLY 1
      RUN ADDY, REPLY 1,REPLY 27,REPLY 99,REPLY,REPLY
      RUN COPY, REPLY 99,REPLY 1,REPLY 105,REPLY )

CALL MAC57
IF REQUEST 'TO FIND KIDNEY ROI''S, TYPE "Y" FOR ISOCONTOURS; "N"
" FOR BUG, METHOD'='Y' THEN CALL MAC44 ELSE CALL MAC50

CALL MAC58
IF REQUEST 'DO YOU NEED TO TRY AGAIN TO DETERMINE THE KIDNEY
"ROI''S?'='Y' THEN CALL MAC42
BREAK MAC12
```

FIG. 4. A literal translation of the program given in Figures 2 and 3 into CL1.

```
( NAME: MAC12 ; subroutine 10
      IF REQUEST 'DO YOU WISH TO MASK THE BLADDER'='Y' THEN
      ( LISTFILE=104, FIRSTFRAME=1, LASTFRAME=27, SUMFRAME=99
         TELZ('Y',1,4),MASK(1),ADDY,COPY(LISTFILE+1)     )
      IF REQUEST 'TO FIND KIDNEY ROI''S TYPE "Y" FOR ISOCONTOURS; "N"
      " FOR BUG, METHOD'='Y' THEN MAC44 ELSE MAC50, MAC58
      IF REQUEST 'DO YOU NEED TO TRY AGAIN TO DETERMINE THE KIDNEY
      "ROI''S?'='Y' THEN MAC42    )
```

FIG. 5. A more elegant translation of the same program.

6. CONCLUSIONS

Initial experience has shown that the use of a command language is likely to be profitable. Three separate areas were considered, being:-

1) the structure of the system

2) the command language

and 3) design of program modules.

Clearly this is an ongoing subject and no single worker could or should present an all-embracing implementation. It is hoped that some of the ideas presented here will be picked up by an appropriate task group and provide the basis for clinically useful systems.

ACKNOWLEDGEMENTS

The authors would like to thank numerous colleagues for informative discussions on these topics and particularly: W. McClain, R. Henne, R. Brill, J. Ericson, J. Gardener, K. Roberts, F. Soussaline and many others.

Part of this work was supported under research contract A/C 334/B94/13.

REFERENCES

1. A. Toffler, Future Shock, Bodley Head, 1970.

2. "Evaluation of the Neonatal Application Package", UCH Internal Report (1977).

3. B.W. Kernighan and P.J. Plauger, Software Tools, Addison/Wesley, Reading, Mass. 1976.

4. O.J. Dahl, E.W. Digkstra and C.A.R. Hoare, Structured Programming, Academic Press, 1972.

5. A. Todd-Pokropek, "The Use of Macro Functions", Informatek Users Group Conference, Clermont Ferand (1977).

6. D. Simpson, Job Control Languages - Past, Present and Future, NCC Publications, ISBN 0 85012 1191, 1974.

APPENDIX 1 - THE 'FORMAL' SYNTAX

While it is probably not useful to give a full BNF definition of the syntax, it is felt that some attempt should be made to describe the language formally. The description which follows adopts the general form of BNF but is not exhaustive and does not adhere to the strict conventions.

Key Symbols

( )

NAME:

IF UNLESS THEN ELSE REPEAT WHILE UNTIL ON

LOOP BREAK RESTART

PAUSE RESUME CALL RUN STOP

ASK REQUEST GET REPLY PRINT SYNCH

Operators

assignment operator: =

arithmetic operators: (infixed) + - * /
(prefixed): -

relational operators: = /= < > <= >= (EQ NE LT GT LE GE)

logical operators (infixed): AND (&) OR
(prefixed): NOT

Names

A name is a sequence of the characters A-Z and 0-9 including no other or newline. The first character of a name must be a letter. Module names must obey the host system conventions.

Literals

Literals are numbers e.g. 1, 1.5, 365.867 or any text enclosed in single quotes ('). Single quote may only appear within literal text if immediately preceded by another. Text literals may only occupy multiple lines if the second and subsequent lines start with (") which as above may be used doubly as an escape.

The Syntax

program	= <named block>
named block	= (NAME:blockname<block body>)
blockname	= name
un-named block	= (<block body>)
block	= <named block> or <un-named block>

```
block body           = <statement> or <statement><block body>
statement            = <block> or <assignment statement> or
                       <conditional clause> or <repeat clause> or
                       <on clause> or <transaction clause> or
                       <command clause> or <expression>
assignment statement = name, assignment operator <expression>
expression           = name or <operation> or literal or (<expression>) or
                       ASK [literal] or
                       REQUEST [literal]
operation            = prefixed operator<expression> or
                       <expression>infixed operator<expression>
logical expression   = prefixed logical operator <logical expression> or
                      <logical expression>infixed logical operator<logical
                       expression> or <expression>relational operator
                       <expression>
conditional clause   = IF or UNLESS,<logical expression>THEN<statement>
                                                   [ELSE<statement>]
repeat clause        = [WHILE or UNTIL,<logical expression>]REPEAT<statement>
on clause            = ON<logical expression>THEN<statement>
transaction clause   = REPLY<expression> or
                       GET name or
                       PRINT<expression>
command clause       = [CALL]<blockname> or
                       [RUN] module name or
                       PAUSE or RESUME or STOP or
                       LOOP[<blockname>]
                       BREAK [<blockname>]
                       RESTART [<blockname>]
                       SYNCH name
```

THE DESIGN AND EVALUATION OF A COMMAND PROCESSING SYSTEM FOR SCINTIGRAPHIC IMAGE ANALYSIS

Bruce R. Line(1), Gerald S. Johnston(2), and James J. Bailey (1)

(1) Laboratory of Applied Studies, Division of Computer Research and Technology;
(2) Department of Nuclear Medicine, Clinical Center;

National Institutes of Health, Bethesda, Maryland, 20014

ABSTRACT

To computer process scintigraphic images, many decisions are made by an operator to control the sequence of analysis and image operations. When the decisions are based on the state of the data, the control requirements may be satisfied using a list of commands incorporating data tests; thereby reducing technical time and effort.

To determine if large clinical analyses could be controlled satisfactorily using automatic instruction interpretation and data testing, a command processing program was written for a minicomputer. The program was designed to: (1) accept commands either from an operator or from a file of instructions; (2) control the order of data processing; and (3) automatically execute large sequences of instructions. The program uses a simply structured command syntax with basic control operations (conditional branching, subroutine calls, and multiple levels of nested sequence iteration) which are adequate to guide most commonly encountered scintigraphic analyses.

Comparisons of the analysis time and error rate between automated command processing and interactive command entry illustrate the advantages of this system in the execution of large scintigraphic analysis procedures.

INTRODUCTION

Specialized computer systems are employed to process scintigraphic data to extend diagnostic accuracy and study organ function. To produce a computer analysis, an operator usually interacts directly with the processor by entering series of instruction steps (1-12). Significant time and effort, however, is required to process clinical protocols of more than 100 steps. In addition, fatique and boredom may increase the frequency of errors in analyses of this size. Fortunately, in most of the interactions, relatively simple decisions are made which are based on the state of the data under analysis. A command management algorithm including processing control structures often will meet these decision making requirements and result in substantial savings in time and energy.

To test this assertion a command processing program was developed which (1) accepts a description of the analysis task in terms of three letter mnemonics and arguments; (2) contains processing control structures (conditionals, subroutine calls, sequence and iteration control); and (3) automatically executes large sequences of commands retrieved from stored lists. To evaluate the utility of this approach, processing time and error rate comparisons were made between automatic and interactive processing of command sequences in a large clinical protocol.

DESCRIPTION OF PROGRAM

The command processing program is written in FORTRAN II and manages analysis subprograms of eight thousand words or less. The program (Figure 1) consists of two sections: the instruction scanner which reads and decomposes instruction sets into commands and arguments; and the function manager which executes the subprograms necessary to carry out the instructions. This simple structure is sufficient to (1) independently process long sequences of commands (automatic mode); (2) process special studies step by step (interactive mode); and (3) maintain independence of the command processing program and the subprogram coding thereby simplifying system maintenance and the addition of new analysis functions.

Instruction Scanner

All communication with the analysis functions is processed by the instruction scanner which consists of three parts: (1) a reader routine which acquires the command strings from the input channel and checks for special characters such as rubout, backspace, escape, and end-of-line markers; (2) a syntax decoder which breaks the instruction text into commands and arguments; and (3) symbol tables which contain definitions of commands and user terms.

Commands are read from either files of instructions or from devices such as the computer console. After input has been collected and checked by the reader the instruction decoder examines each character using format conventions to identify commands, arguments, and comment strings. Commands and arguments are stored in sequence in separate stacks and are later retrieved by the function manager during command execution.

The instruction conventions were chosen because of their simplicity, and in general, may be summarized by one rule: "A command is followed by its arguments." The general flow of instruction processing is shown in Figure 2.

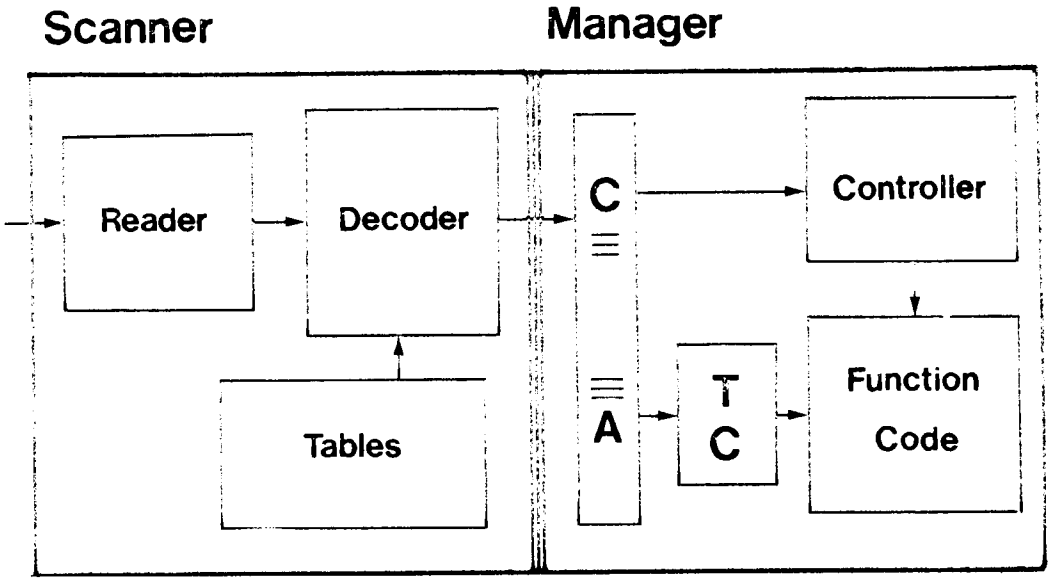


Figure 1. System block diagram. Input from an instruction file or device is checked and decoded using tables of command and argument definitions. The command stack (C) is used by the controller to determine the order of function requests. The type conversion routines (TC) are used to retrieve the arguments from the argument stack (A).

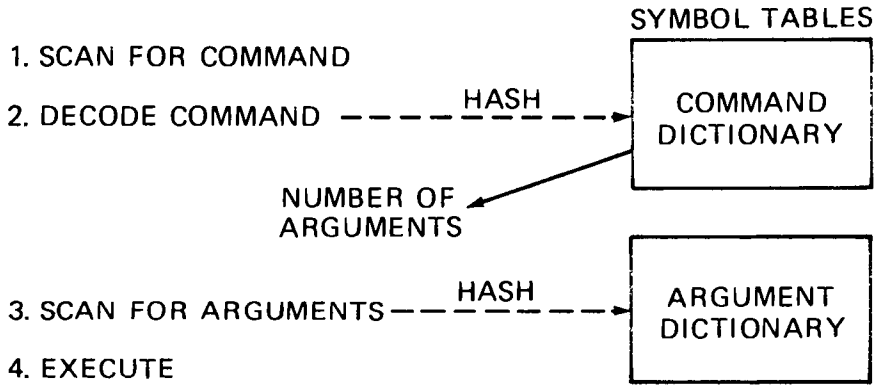


FIGURE 2. Flow of instruction processing.

Commands

Commands are represented with strings of alphabetic characters and are located either after an "escape" character or when the arguments to the previous command have been satisfied. The command symbol table, or dictionary, is used to identify the requested operations. This dictionary contains the location of the program coding and the number of arguments required by each computer function. It is referenced using a hashing algorithm (13) to produce a table entry from simple numeric operations on the first three non-blank letters of the command name. The letters "INP" are thus sufficient to identify the "INPUT" command in the following instruction which changes the input device to the paper tape reader:

INPUT = :TAPE.READER.

Arguments

An argument may be a text string or a number having either real, integer, or octal data type. In addition, symbolic arguments may be used for numbers to clarify protocol procedures. Symbols consist of a colon followed by a string of letters, digits, or the character ".". A symbol definition table for numeric values is referenced whenever a symbol identifier ":" is encountered in the instruction text. For example, the following symbols:

```
:123
:APRIL.1973
:BLOOD.TYPE..APOS,
```

refer to numbers, the values of which are determined from the definition table. All numeric arguments are temporarily stored as real values in the argument stack and are converted to the data type required during function execution.

In the example command shown above, the command dictionary indicates that the "INPUT" function requires one argument so the line is scanned until the number ":TAPE.READER" is located.

In applications requiring the use of variables, arguments are accessed by referencing memory. Instead of representing the value of an argument, a numeric term or symbol represents the position of the argument in a storage area for variables. Whenever symbols or numeric strings are followed by a "! " mark they are interpreted as the locations of integer values, whereas a "%" suffix is interpreted as the location of a real value. For example,

```
:NEXT.MAGTAPE.FILE!,
```

references a integer value which is used to represent the number of the next magnetic tape file.

Comments

Any text which does not satisfy the criteria for a command or an argument is considered comment. In situations where there may be confusion, comment text is enclosed in single quotes.

Function Manager

Automatic processing

The function manager retrieves commands from the command stack in the order of their arrangement in the instruction list. The stack contains information which the function manager uses to prepare the computer for execution of an analysis subprogram. The subprogram requests any required arguments through type conversion routines which change the real values in the argument stack to integer or octal data type, when necessary, or retrieve a string argument from storage.

If either of the stacks is found to be empty when a request is made, the instruction scanner is restarted to process more command text. Since the scanner may be directed by the command syntax to retrieve arguments from memory locations, the results of previous operations can be used in the control of later functions.

Interactive processing modes

The interactive mode of operation may be employed by an operator in the design of an automatically processed procedure or to execute simple analyses which arise when close examination of data is indicated. When the command processor is executing instructions interactively, and no arguments are supplied on the line containing the command, the system identifies the function requested and prompts the operator to supply the necessary arguments (Figure 3).

The prompting interactive mode of command entry is often useful when the sequence of required arguments is in doubt. No prompting occurs (terse mode) when the command and any of its arguments are supplied on the same line. Users familiar with the command arguments are thus not forced to wait for output intended to aid novice users.

Control Functions

Processing control commands are provided to extend instruction file versatility (Figure 4). Line numbers are used by the function manager to transfer control both within and between instruction files. The "within" file control commands are GOTO, IF, and DO. The GOTO command takes a line number as an argument and provides an unconditional transfer to that point in the active file. For example, the command "GOTO 10" forces a jump to the instructions on line 10. Also legal are statements like:

```
GOTO :LAST.LINE
GOTO :COMPUTED.ADDRESS! ,
```

which have symbolic arguments or reference line number values which may be the result of a computation.

Processing control may be transferred conditionally with the IF statement which tests a numeric value by comparing it to zero. Control is passed to a line depending on whether the tested value is less than, equal to, or greater than zero. For example, the command:

```
IF(:VALUE!):COMPUTED.ADDRESS!,:EXIT,:ERROR.ROUTINE,
```

tests the contents of the memory location designated by the symbol ":VALUE" and begins processing the instructions on the line number ":COMPUTED.ADDRESS!" if that value is less than zero; on line ":EXIT" if equal to zero; and on line ":ERROR.ROUTINE" if greater than zero.

The DO statement provides a loop structure, which may be nested to a depth of 10 levels. This statement contains an end of loop line number, an index, and three index control terms. The command:

```
DO TO :LAST.LINE, :INDEX FROM :START! TO :END! BY:INCREMENT!,
```

starts an execution loop with an endpoint specified by the symbol ":LAST.LINE". The loop index is the value in the storage word referenced by ":INDEX". The variables ":START, :END, and :INCREMENT designate the first, last and increment index control values respectively.

Transfer of control between files is possible using either the "BATCH" command which passes control to the first instruction line of the target file, or the "BEGIN" command which may be used to begin execution at any line of the requested file. For example, the command:

```
BATCH :HEART.EJECTION.FRACTION,
```

PROMPTING MODE

User Command:	RMT
Command Identification:	READ MAG TAPE FILES
System Prompt:	FILE SPAN, IMAGE SPAN, TIME FUNCTION SPAN?
User Input: [Execution . .]	1 , 1 1 , 5 1 , 5

TERSE MODE

User Command:	RMT 1, 1 1 , 5 1 , 5

FIGURE 3. Interactive processing modes.

TRANSFER	GOTO 10, GOTO :COMPUTED.ADDRESS!
BRANCH	IF (:VALUE!) :ADDRESS!, :EXIT, 125
ITERATION	DO TO :LAST.LOOP.LINE! :INDEX FROM :START! TO :END! BY :INCREMENT
SUBROUTINE CALLS	BATCH :HEART.EJECTION.FRACTION BEGIN :UTILITY.FILE, AT LINE # 130

FIGURE 4. Example of processing control functions.

executes the file ":HEART.EJECTION.FRACTION" and the command:

BEGIN :UTILITY.FILE AT :READ.ROUTINE,

begins the ":UTILITY.FILE" at the line number designated by the symbol ":READ.ROUTINE". The terminal statement in any file or file portion must be an END statement which signals the system to return control to either the calling instruction file or to the interactive processing mode if no calling files exist. Although only one file is considered active at a time, several files may be in different stages of completion during the execution of an analysis procedure. The function manager keeps a stack of values to determine the stage of completion of each file and the sequence of their calls.

This system has been designed to permit well defined sequences of commands to be processed without operator intervention. However, processing pointers are maintained so that execution may be stopped between instructions to examine intermediate results, make corrections if necessary, and then restart the analysis without loss of continuity. Figure 5 shows a typical

```
1        'COMMAND LIST FOR CREATING FUNCTIONAL MAPS OF . . .
2   LEFT VENTRICULAR PEAK EJECTION RATE AND TIME TO PEAK EJECTION'

3   'STORAGE REGISTERS 41 THRU 53 ARE ASSIGNED NAMES'
4          "F"      < 41      'FIRST INPUT FRAME NUMBER'
5          "L"      < 42      'LAST INPUT FRAME NUMBER'
6          "FS"     < 43      'FIRST SUMMED FRAME NUMBER'
7          "LS"     < 44      'LAST SUMMED FRAME NUMBER'
8          "LSM2"   < 45      'LAST DIFFERENCE FRAME (FOR LOOP CONTROL)'
9          "D"      < 46      'CURRENT DIFFERENCE FRAME (LOOP CONTROL)'
10         "S"      < 47      'CURRENT SUMMED FRAME      (LOOP CONTROL)'
11         "SP2"    < 48      'CURRENT SUMMED FRAME + 2  (LOOP CONTROL)'
12         "X1"     < 49      'FIRST DIFFERENCE FRAME IN CURRENT SET OF 3'
13         "X2"     < 50      'SECOND DIFFERENCE FRAME . . .'
14         "X3"     < 51      'THIRD DIFFERENCE FRAME . . .'
15         "XR"     < 52      'RESULTING SMOOTHED FRAME'
16         "DM1"    < 53      'LAST DIFFERENCE FRAME – 1'
17  'PREFIX LETTERS SPECIFY DATA TYPES: F = FRAME, S = STORAGE LOCATION
18  SUFFIX CHARACTER ! SPECIFIES INDIRECT REFERENCE'

19  'CLINICIAN SPECIFIES DESIRED FRAME SEQUENCE'
20         TYPE "ENTER FRAME SPAN"
30         ISI :F, :L
40  'SUM SETS OF 3 FRAMES EACH. STORE RESULTS IN FRAMES 1 AND UP'
41         SUM 3 FROM FRAME:FI TO :LI RESULTS BEGIN AT FRAME 1
50  'COMPUTE NUMBER OF LAST SUMMED FRAME'
51         (S:F – S:L + 1) / 3 = S:LS
52  'COMPUTE NUMBER OF FIRST SUMMED FRAME (=1)'
53         ZERO + 1 = S:FS
60  'SMOOTH THE SUMMED FRAMES.  OUTPUT RESULTS TO FRAMES 1 AND UP'
62         SMF FRAMES:FS!  THRU  :LS!  RESULTS BEGIN IN FRAME 1

70  'CREATE DIFFERENCE FRAMES TO SHOW CHANGES IN VOLUME . . .
72  AND OUTPUT THE RESULTS TO FRAMES 101 AND UP'
80         S:LS – 1 = S:LSM2
81         DO THRU LINE 90 FOR S:S=:FS!  TO :LSM2!  STEPPING BY 1
82         S:S + 100 = S:D
83         S:S + 2 = S:SP2
84         'SUBTRACT.  THEN ZERO OUT NEGATIVE ELEMENTS'
90         F:SP2! – F:S!  THRESHOLD 0 = F:D!

100 'SMOOTH DIFF FRAMES.  OUTPUT RESULTS TO FRAMES 202 AND UP'
109        S:D – 1 = S:DM1
110        DO THRU LINE 120 FOR S:X2=F102 TO S:DM1 STEPPING BY 1
111        S:X2 + 1 = S:X3 – 2 = S:X1 + 101 = S:XR
120        F:X2! * 2 + F:X1! + F:X3! / 4 = F:XR!

130 'TRANSFER SMOOTHED FRAMES BACK TO FRAMES 101 AND UP'
131        TRANSFER FRAMES 201 THRU :XR!  TO FRAMES 101

140 'MAKE FUNCTIONAL MAP OF PEAK EJECTION RATE'
141        MXM FRAMES 101 THRU :DM1, AND RESULT MAP IN FRAME 100

150 'MAKE FUNCTIONAL MAP OF TIME TO PEAK EJECTION'
151        EQM REFERENCE FRAME 100, SEQUENCE FRAMES 101 THRU :DM1!
152                AND STORE MAP IN FRAME 99
153 END
```

FIGURE 5. A typical instruction file for automatic analysis.

analysis file which demonstrates the appearance of the instruction set used to produce functional maps of the left ventricular peak ejection rate and the time to peak ejection.

Clinical Protocol for Scintigraphic Ventilation-Perfusion Analysis

The analysis of pulmonary ventilation-perfusion studies can include

generation of images of ventilatory flow, ventilation-perfusion ratios, and estimates of regional gas exchange (14). An automatically processed clinical protocol has been derived to produce these images as part of a ventilation and perfusion lung study. This protocol has been developed and tested using a data base of 550 patient studies stored on magnetic tape. The procedure is executed from several files of instructions which contain a total of 630 function requests and 970 arguments. Each command actuates a subprogram from a set of over 200 operations presently contained in the function library.

Comparison of Automatic and Interactive Processing Modes

One hundred sequential patient studies were selected and each was inspected for errors related to technical procedure, patient cooperation, and equipment malfunction. Of this group five studies had errors due to patient movement, four had suboptimal count rates during the ventilation study, and two contained errors attributable to equipment malfunction. Each of the 100 patient scans was processed in automatic mode; the data was concentrated on four magnetic tapes. Each study contained a perfusion image and up to 155 five second Xenon-127 or Xenon-133 dynamic ventilation distribution images.

Three individuals interactively operated the instruction protocol used to process the ventilation-perfusion analysis. One of the individuals (BL) had extensive experience with the command processing system and the ventilation perfusion analysis. One individual (CQ) had experience with the scintigraphic analysis but little background using the command processor and the third interactive analysis was performed by a medical secretary (LA) who had never processed a scintigraphic study. The instruction set used in the interactive sessions was extensively tested for errors in logic and was shown to produce accurate results if followed exactly. The instruction set used differed from the automatically processed version in the following ways: (1) no line numbers were included; and (2) any branching or loop operations which could not be removed were placed in separate instruction files which were run automatically at the appropriate time in the interactive analysis. The resulting instruction set contained 510 function requests and 830 arguments.

A single patient study was chosen at random from the set of studies processed automatically and was confirmed to be technically satisfactory. This study was preanalyzed and was shown to produce identical results by both the interactive and the automatic mode of analysis.

During interactive analysis, when an error in command entry was identified by the system or by an operator, it was corrected by retyping the instruction. However, if an error was made which was not detected until a later point in the analysis, an error was recorded, and the analysis was corrected up to the point where the error was detected. The errors requiring correction and the processing time for each operator were recorded. When the analysis was finished, it was judged complete if the image results matched the carefully preanalyzed results. The time required to complete 100 studies was computed for each individual assuming error free analyses and a 40 hour work week.

RESULTS

Of the sets of results produced from the automatically processed studies, 94 were considered adequate for clinical interpretation. Two of the

studies were rejected because of patient movement, two had too few counts in the ventilation images, and both of the studies containing errors due to equipment malfunction were not recoverable. The processing of the four tapes was completed over a two day period and consumed 21.30 hours of processing time. The average time to process one study was 12.9 minutes. No errors were attributable to either hardware or software malfunction in the automatic processing of the patient set. The study selected for interactive analysis took 12.2 minutes to process automatically.

The individual results for the interactive analyses (Table I.) showed a mean processing time of 137.7 minutes with an average of 5.3 errors requiring partial reanalysis. None of the interactive analyses were completed correctly, however. In all cases there were differences between the "known" solution and the final results obtained by the interactive analysis.

The mean error rate for the individual analyses was 1.04 per hundred commands and the mean time (typing and execution) between commands was 16.2 seconds. The expected time required to process 100 patient studies, assuming no errors and a 40 hour work week, was 5.7 weeks.

DISCUSSION

The computer analysis of scintigraphic data includes three general phases: (1) collection of detector events; (2) reordering the data into images or time-activity curves and (3) presentation of the results for clinical interpretation. Although all phases could be achieved in one algorithm, limitations quickly arise with respect to program size and complexity. An alternative approach to problem solving involves the use of a library of programs which manipulate data in standard formats. This technique allows a solution to a scintigraphic analysis problem to be expressed through prestructured library routines instead of requiring a compiled program which may involve substantial time and effort to produce. It is more flexible because compatibly structured routines may be sequenced in various manners to define solutions. When the analysis procedure changes, the ability to quickly link a new set of programs gives this more modular approach a definite attractiveness.

An analysis protocol may be directly modified to achieve the desired clinical results through interactive manipulation of the instruction file. A clinically satisfactory procedure can be rapidly achieved by this approach in contrast to the tedious and time-consuming process of changing a large compiled program. Freer experimentation with the data is thus possible in situations where it is difficult to define what the exact solution to a problem should be. Since comments may be liberally included in the files, and because files may be called as subprograms, an analysis problem may be cleanly structured and well documented.

A list of instructions defines the analysis process in terms of the sequence of commands in the list and the values of their arguments. Like a program language, the list represents the state of the analysis of a clinical or research protocol but is more concise; each command representing a computer program.

The command processor is logically separate from the coding for the analysis functions. This is desirable, because new functions may be added to the system with a minimum of restructuring. The interfacing programs for

Table 1. Interactive Analysis Time and Error Comparison

Operator	Minicomputer Experience (Yrs)	Protocol Analysis			
		Experience*	Errors	Time	Time/100 Studies**
BL	> 3	3	5	128.5	5.35
CQ	> 2	None	4	148.5	6.19
LA	None	None	6	136.1	5.67
Mean			5.3	137.7	5.74

*Number of interactive analyses performed before study

**Number of weeks to process 100 studies working 40 hrs/week

argument retrieval and type conversion allow the command processor to collect arguments without prior knowledge of their final use or form.

The control structures are patterned after FORTRAN GO, IF, DO, and subroutine statements. In that these structures are sufficient to generate a highly adaptable program in FORTRAN, it is likely that they should be adequate for most processing tasks encountered in scintigraphic analysis. Because the command list is not compiled prior to execution, the command processor is not intended for use in element by element array analysis. Such tasks are meant to be incorporated in compiled functions which may be easily added to the system. The purpose of the command processor was not to replace the high level command languages but to provide a means by which basic functions may be designed independently and then be recombined in an instruction file to achieve the solution to a scintigraphic analysis problem.

The comparisons between the interactive and the automatic analyses illustrate that: (1) automatic processing can be used to execute large well defined analyses with little direct supervision; (2) interactive processing may be very time consuming when large numbers of directly entered commands are involved in protocol analyses; and (3) analyses of protocols requiring more than 100 interactive steps may be expected to contain errors related to operator command entry. This effectively limits the number of interactions with a processing system to less than 100 steps to yield of an acceptable percentage of error free studies.

The command processing program with its basic control structures provides a vehicle through which interactive manipulation and automatic processing may be combined in a single system. The modularity of the analysis software in system libraries may thus be maintained to preserve flexibility in trial runs which require direct inspection of the data. However, when the modularity is no longer needed, it may be suppressed during automatic operation to execute large numbers of clinical studies.

In summary, it appears that large scintigraphic analyses are best constructed in modular systems because of the ease of expression of the solution of the problem in terms of a sequence of modular operations. It is not practical however, to control large sequences of commands interactively because of the time required and the number of errors which stem from command entry. The command processing system described offers an effective method of executing such procedures and is an example of a system which should prove highly useful in clinical and research analyses of large scintigraphic analysis protocols.

ACKNOWLEGEMENTS

The authors wish to express their gratitude to Catherine Quigley and Lanita Atkinson for assistance with the interactive analyses, and to Joseph F. Wilson for his invaluable support in the production of the clinical studies and the automatic analyses.

REFERENCES

1. D. W. Brown, "Digital computer analysis and display of the radioisotope scan," J. Nucl. Med. 5, 802-806 (1964)

2. D. W. Brown, "Digital computer analysis and display of the radionuclide scan," J. Nucl. Med. 7, 740-753 (1966)

3. D. E. Kuhl, and R. Q. Edwards, "Modifying and rearranging scan data under direct observations using a magnetic storage drum, high-speed digital circuitry and CRT display," J. Nucl. Med. 8, 289-290 (1967)

4. E. M. Smith, and A. B. Brill, "Progress with computers in nuclear medicine, Nucleonics 25, 5, 64-71 (1967)

5. S. Kaihara, T. K. Natarajan, and H. N. Wagner, "Construction of a functional image from regional rate constants," J. Nucl. Med. 10, 347 (1969)

6. A. B. Brill, J. P. Erickson, and J. Patton, "Computer programs used in nuclear medicine clinic and research laboratory," In "Sharing of Computer Programs and Technology in Nuclear Medicine, AEC CONF-710425, Oak Ridge, Tenn, 123-146 (1971)

7. R. L. Hill, and J. R. Cox, "A system of programs for a small computer interfaced to a gamma camera," In Sharing of Computer Programs and Technology in Nuclear Medicine, AEC CONF-710425, Oak Ridge, Tenn, 113-118 (1971)

8. W. G. Monahan, "Data analysis for an on line gamma camera system," In Sharing of Computer Programs and Technology in Nuclear Medicine, AEC CONF-710425, Oak Ridge, Tenn, 105-112 (1971)

9. H. N. Wagner, and T. K. Natarajan, "The computer in nuclear medicine, In Sharing of Computer Programs and Technology in Nuclear Medicine, AEC CONF-710425, Oak Ridge, Tenn, 51-60, (1971)

10. T. F. Budinger, "Clinical and research quantitative nuclear medicine system." In Medical Radioisotope Scintigraphy," Vol. 1, Symp., Monte Carlo, 1972, IALA, Vienna, 501-555, 1973a.

11. N. M. Alpert, L. A. Burnham, L. A. Deveau, J. E. Correll, D. A. Chesler, S. M. Pizer, G. L. Brownell, "Numedics: A system for on-line data processing in nuclear medicine," J. Nucl. Med. 16, 386-392 (1975)

12. T. F. Budinger, and J. Harpootlian, "Developments in digital Computer Implementation in Nuclear Medicine Imaging," Comput. Biomed. Res. 8, 26-52 (1975)

13. E. G. Coffman, Jr., and J. Eve, "File structures using hashing functions", CACM, 13, 427-432 (1970)

14. R. G. Crystal, J. D. Fulmer, W. C. Roberts, M. L. Moss, B. R. Line, and H. Y. Renyolds, "Idiopathic pulmonary fibrosis: clinical, histologic, radiographic, physiologic, scintigraphic, cytologic, and biochemical aspects," Ann. Int. Med. 85, 769-788 (1976)

Command Languages

Chairman: Stephen M. Pizer (Chapel Hill, NC, USA)

Command languages are a mechanism to provide the capability of 1) controlling the order in which processing modules are executed, 2) determining the files which are the input and output of the process as well as intermediate data communicating between modules, and 3) specifying and controlling the specification of the parameters of the processes. The two papers in this session described one implemented command language (Line) and one proposed command language based on two previously implemented languages (Todd-Pokropek). Both authors emphasized the following reasons for using a command language.

1. Convenience for physicians and technologists to specify processing in terms of functions familiar to them and in a language using familiar terminology.
2. Flexibility of specifying protocols not originally provided by the manufacturer.
3. Considerable savings in elapsed time and user time over direct specification of each subprocess and all of its parameters by typing or pushing buttons.
4. Convenience of running in batch mode (when the user does not need to be present).
5. Lowered error rate, since the user has to type much less once a protocol is defined.
6. Capability to share protocols among institutions, especially if the command language can be standardized.

Both authors emphasized the need for flow control statements in the command language, e.g. for repetition and conditional selection.

The discussants were encouraged not only to comment on the individual papers but to help define what is an appropriate command language for computing in nuclear medicine. This matter is being studied further by a task group of the Computer Council of the Society of Nuclear Medicine, chaired by Andrew Todd-Pokropek.

Questions on the specific systems included the following. Mark Moore asked Bruce Line why his command language used control structures which have been shown difficult to use, such as GO TO. Stephen Pizer suggested that physicians would find that the FORTRAN constructs used were unnatural and that using these constructs because FORTRAN was the programming language used was a mistake. Bruce Line agreed that some improvements could be made but stated that his system worked, was easily learned, and was used by physicians and that some of the above choices were a matter of taste. Alan Ashare asked whether Line's system was portable to other machines, with the answer that the system was written in FORTRAN II but with an underpinning of nine fairly small drivers in assembler language.

William Winspur asked Andrew Todd-Pokropek why the FORTRAN WRITE command was not included in his proposed language. The reply was that the user should not have to do his own buffering

and yet needed to be able to specify more things than WRITE allowed, e.g. position on a TV screen. That is, a more flexible subroutine to do writing needed to be provided as a utility. Winspur asked how synchronization would occur with pipelines, and Todd-Pokropek answered that this should be handled by the operating system and that the use of the pipeline should be transparent to the user writing or executing protocols.

Some of the discussion addressed the issue of the speed losses due to overhead of interpretation or subroutine calls when using a command language, and the possibility of compilation and optimization either automatically or by rewriting protocols using assembler language. Both authors pointed out that, though there was some overhead when using a command language, most of the time was spent executing the program modules and only a negligible fraction was spent interpreting commands. In fact, Todd-Pokropek believed the command language should have no direct arithmetic capability. The need was for the modules to execute fast and thus when commonly used, they needed to be written in assembler language or compiled with an optimizing compiler. Neither author had run benchmarks for computation speed. But both argued that in terms of elapsed time, the time saved over fully interactive use of programs provided by manufacturers was very great.

The most spirited discussion was on the need for and desirability of command languages. Many wondered why a subset of a language that was already provided by computer manufacturers would not be a suitable command language, as it had control structures and subroutine capability. Languages suggested included FORTRAN, FOCAL, and APL. It was pointed out that some of these languages did not do file management or flexible input/output or handle core overlays, and these languages were hard for the physician to use as they did not use familiar terms. Mark Moore pointed out that Todd-Pokropek's proposed language was not specific to nuclear medicine but was a more general purpose process control language. He wondered why computer scientists shouldn't create the language, given that careful planning was necessary to avoid another bad language like FORTRAN. Andrew Todd-Pokropek stated that this would be fine, but no such language now existed, whereupon Mark Moore warned that this might be because of the difficulty of producing such a language.

Michael Goris argued that the fact that most of the work done for this meeting had not needed systems with command languages indicated they were not needed, especially since they would often fail to contain modules which did new things needed in research. Both speakers argued that if the modules were atomic enough, they could be used as building blocks in research and in any case they would do the overhead of the research (handling files, input/output, etc.), leaving the meat for the researcher to put in a new module. But both pointed out that the biggest payoff with command languages came in routine clinical use rather than in research.

Jack Goldberg strongly questioned the need for command languages and, representing a manufacturer opposed an attempt to make manufacturers support such a language. He argued that only buttons, augmented by interaction of po[illegible] with a cursor and

entering parameter values when necessary, provided a proper user interface for routine clinical use. When a new protocol needed to be defined, he said, it could be remembered by the system from a series of button pushes selected from a menu, or a new program could be written in a high or low level language of the computer and "put under" a spare button. He felt that physicians should not be forced to learn a new language and that programming should be done by a computer person. Both speakers and others argued that flow control commands were essential to appropriate flexibility and time saving for the physician. They stated that a command language was not incompatible with a system based on button pushing, but rather it allowed button pushes to cause protocols in command language to be executed, i.e. the command language to be transparent when it was not necessary. They said it was possible for new users to learn the command language progressively, while using it.

Nicholas Brown pointed out that some physicians did not like to type, whereas Alan Ashare expressed concern that command languages might make it too easy for physicians to create protocols when they didn't understand the effect of the modules. Keith Britton supported the usefulness of command languages based on his experience with one produced by Andrew Todd-Pokropek, but he stated that computer scientists together with physicians should create new protocols. Stephen Pizer suggested that such teams should create large protocols, but the command language should be available to the physicians to create short ones conveniently. Michael Tobin said that in their experience with one of Todd-Pokropek's command languages, people unable to program in FORTRAN could easily create protocols and that the resulting routines were reliable and fast. Michael Weisner said command languages in their hands allowed the quick creation of new protocols and allowed recovery from studies which had been fouled up. Andrew Todd-Pokropek repeated the finding that though command languages were useful in research and special cases, they were most useful in clinical routine, where it turned out that fairly long protocols were common, e.g. with 300 statements, largely consisting of parameter specifications and control statements and containing only a few commands to run modules.

Keith Britton said command languages, their processors, and the associated modules had to be frozen if they were to be clinically useful. Data collected from a routine system could be brought to research computers for new protocols. He noted unhappily but understandingly that the final stages of debugging of the language processor, modules, and protocols might have to involve physicians, but he said that this stage had to be passed quickly, since an early freeze was essential to allow long term clinical use. Andrew Todd-Pokropek pointed out that with small modules, a system could be robust and allow protocols and modules to be frozen while adding other protocols and modules to allow flexibility, which is required. It was noted that freezing the command language and modules was also necessary if the command language and programs in it were to be portable.

Michael Weisner said that it would be difficult to define a language processor and modules which would fit in all machines. He and Andrew Todd-Pokropek agreed that a debugging trace was very

useful. They also agreed that how to deal with files was a difficult problem, but both Andrew Todd-Pokropek and Bruce Line answered, to a question by Leslie Rogers, that the file handling facilities of the operating system formed the basis of and the constraints on the file handling done by the utilities. These utilities are system dependent programs which carry out the file handling needs of the modules, which are not system dependent and are thus portable like protocols in command language.

FACTOR ANALYSIS OF SEQUENCE SCINTIGRAMS

P.Schmidlin
Institute of Nuclear Medicine
German Cancer Research Center
Heidelberg, Germany

ABSTRACT

Factor analysis was applied to isotope nephrograms taken with 131-I-hippurate and with 99m - Tc - PTT. A Karhunen-Loève expansion was performed on typical time-activity curves. The evaluation of nephrograms was performed by calculation of the contributions of the Karhunen-Loève-functions, the so-called factors, to the individual time-activity functions. These factor contributions were displayed like scintigrams on a color-TV. Mean factors over the whole kidney area were taken as a measure for the quality of the kidney. The evaluation of the hippurate studies gave precise results, whereas the results of the PTT studies were not yet quite satisfactory.

Mathematical Principles

Factor analysis is used for feature extraction in pattern recognition. It enables to decide between important and unimportant features and to look for common properties of different patterns. In this paper a version of factor analysis is described which calculates the principle components of time-activity curves. The method is applied to isotope nephrograms. First results have been reported in 1976 (1).

An isotope nephrogram may be considered as a sequence of count rate pictures. The count rates are described by

$$n(x,t) \qquad x = 1....X,\ t = 1...T \tag{1}$$

where x is the localization (normally a two-dimensional vector) and t is the time. The sums over all time slices and the sums over all cells in a time slice are given by

$$n(x) = \sum_{t}^{T} n(x,t) \qquad n(t) = \sum_{x}^{X} n(x,t) \tag{2}$$

The total count rate in the study is

$$n = \sum_{x,t} n(x,t) \tag{3}$$

The mean values of the count rates over all times and over all picture cells are given by

$$\bar{n}(x) = n(x)/T \qquad \bar{n}(t) = n(t)/X \tag{4}$$

The time-activity curves in different picture cells can be correlated. This is described by the sum of the deviation products

$$S_{tt'} = \sum_{x} (n(x,t)-\bar{n}(t))(n(x,t')-\bar{n}(t')) \tag{5}$$

From this the correlation matrix is derived:

$$s_{tt'} = S_{tt'} / \sqrt{S_{tt} \cdot S_{t't'}} \tag{6}$$

In this paper a different algorithm established by BENZECRI (2) is used. Instead of the deviations from the mean value the deviations from zero are used:

$$S_{tt'} = \sum_{x} (n(x,t) . n(x,t')) \tag{7}$$

Then the correlation matrix is

$$s_{tt'} = \qquad / \sqrt{n(x)n(t) \cdot n(x)n(t')} \tag{8}$$

This correlation matrix is inserted into an eigenvalue equation:

$$\sum_{t'} s_{tt'} G_k(t') = \lambda_k G_k(t) \tag{9}$$

The λ_k are the eigenvalues of the correlation matrix s. They are calculated by diagonalization of this matrix. The G_k are the eigenfunctions of the equation. In the case of factor analysis they are called the factors. The G's form a complete orthogonal system of fundamental functions. Therefore any time-activity curve may be reconstructed by means of these factors:

$$n(x,t) = \frac{n(x)n(t)}{n} \sum_{k} \frac{F_k(x)G_k(t)}{\sqrt{\lambda_k}} \tag{10}$$

In this equation F_k are coefficients which determine the contribution of a certain factor to a time-activity curve. They are called factor loadings. The expansion described in the previous equation is called a KARHUNEN-LOEVE expansion. The coefficients F_k of the expansion are calculated by means of the orthogonality relations

$$F_k(x) = \frac{1}{\sqrt{\lambda_k}} \cdot \frac{1}{n(x)} \cdot \sum_{t} G_k(t)n(x,t) \tag{11}$$

Any time-activity function may be characterized by the calculation of the coefficients F_k.

An important feature of this algorithm is the compression of the information in very few factors. In the evaluations of kidney functions already the first factor is a very sensitive measure of the quality of a kidney. The higher factors beginning with the third or fourth factor, do not contain any essential information, but they describe almost random fluctuations.

There is some difference between factor analysis and KARHUNEN-LOEVE-expansion (3). Factor analysis originally decides between common (correlated) properties and individual (uncorrelated) properties. The expansion takes into account both kinds of properties. The disadvantage is that there are more unknowns than equations in this case. So the factors as well as the factor loadings are not uniquely determined. In KARHUNEN-LOEVE expansion the higher factors may be considered as the individual properties of the objects. In this case there is no principal distinction between both methods, but in KARHUNEN-LOEVE expansion factors are uniquely determined and there are no arbitrary rotations of the coordinate system.

The evaluation of an isotope nephrogram could result in the extraction of factors from the correlation matrix and then the calculation of factor loadings. In this case the factors would have a different meaning for every kidney study. Because the factors are a fundamental function system, any time activity curve can be expanded by means of the same factors. This way has been chosen, because it makes results more meaningful.

Thus the evaluation of sequence scintigrams consists in two steps: first, a set of good reference factors is calculated for every type of study. The second step is the evaluation of the individual sequence scintigram by calculation of the factor loadings.

Calculation of reference factors

To get a valuable set of reference factors, the correlation matrix of different time-activity curves is diagonalized and the eigenfunctions are calculated. Different ways of getting representative collectives of time-activity curves were tested. The sampling size of the picture frames was 8 x 8 cells. The calculations were performed on an IBM 370-158 computer.

A set of 48 picture cells was selected from six different kidney studies to get a factor set for the evaluation of kidney function studies with ^{131}J-hippurate. These cells were selected by hand to get time activity curves as different as possible from each other (1).

For the evaluation of perfusion studies with ^{99}Tc-pertechnetate a collective of time activity curves has been extracted from one perfusion study (1). This factor set was unsuccessful in practice and has been dropped.

To improve the selection of a good collective of reference cells, a learning program was established. The flow diagram of this program is shown in fig.1. The procedure begins with the manual selection of a collective of four cells of a kidney study. From the time activity curves of these four cells, G-factors are calculated. Then another kidney study is fed into the procedure.

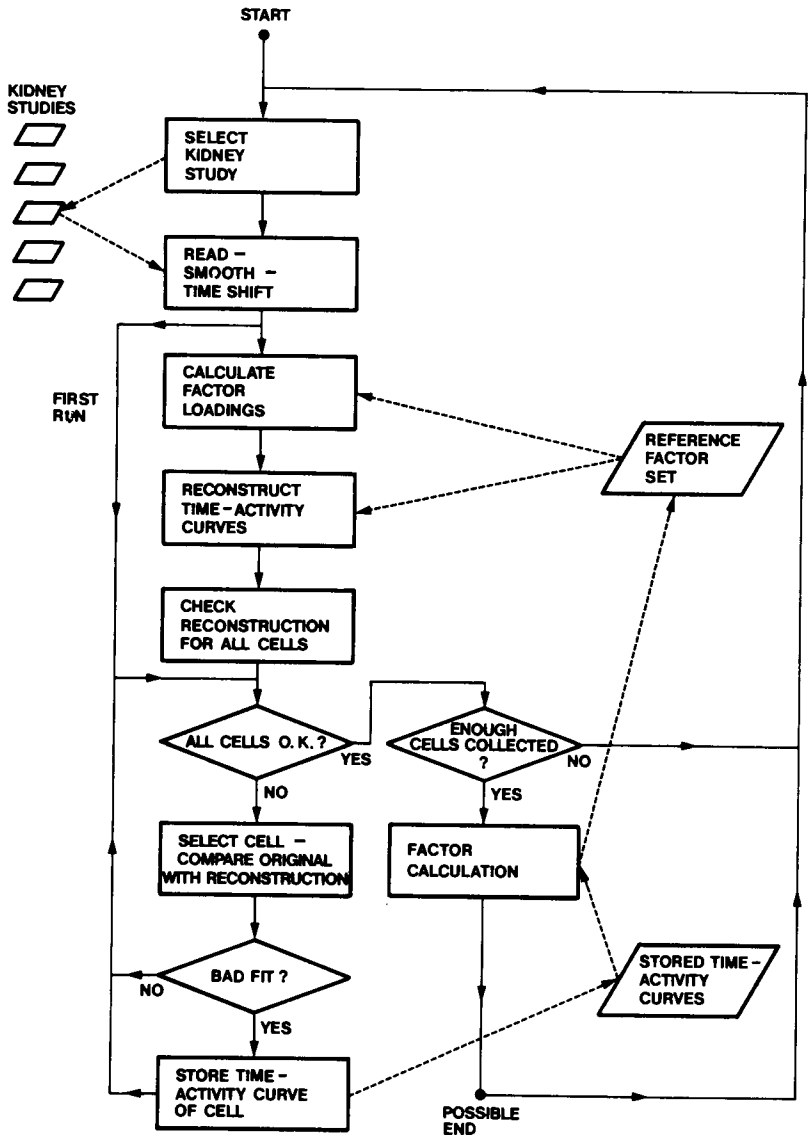


Fig.1: Flow diagram of the learning program for factor calculation

The time activity curves of all 8 x 8 cells of this study are reconstructed by means of the previously calculated G-factors. The quality of the reconstruction is taken as a measure for the quality of the previously calculated G-factors. The reconstruction quality is tested by means of the sum of the deviation squares. The four cells which show the biggest deviations are added to the cell collective. From the resulting eight time activity curves new G-factors are calculated. Then other studies are added, and other badly reconstructed cells are added to the collective. Thus G-factors are calculated for collectives of about 16, 32 and finally 64 cells.

This iterative procedure was applied to the pertechnetate perfusion studies. 64 picture cells were extracted from 23 different perfusion studies. Finally all perfusion studies could be reconstructed by means of the resulting G-factors with sufficient accuracy. Fig.2 shows the 3 biggest factors calculated by this procedure.

In order to avoid the development of a new factor set for each kind of study, it was tried to establish theoretical time activity curves, which would describe any kind of accumulation and excretion studies. The extracted factors for the evaluation of the hippurane-studies were rather bad. The method failed completely to establish factors for perfusion studies. This result leads to an important conclusion: The method is most sensitive, if the inserted time activity curves are most specific for the type of study to be evaluated. This means, that e.g. one kind of study should not be evaluated by means of the factors extracted from another type of study. Furthermore the calculation of factor sets by means of theoretical considerations or models is not sufficient for practical purposes.

Evaluation of isotope nephrograms

The individual kidney study is evaluated by calculating the contributions of the factors to the time activity curves, i.e. the factor loadings. It may be performed in two different ways: either mean factor loadings may be computed over regions of interest. These mean factor loadings provide a good measure for the kidney function. The other possibility is the calculation of the factor loadings of all picture cells. The two-dimensional grey scale display of the loadings of the first factor would represent already a valuable factor scintigram. Using a color display, it is possible to assign the three colors to the first three factors to get one factor scintigram which contains the information of the first three factors. This is performed in our lab by means of a commercial color TV which has been modified and connected to a CAMAC-Interface. The information of a dynamic study is represented in one 64x64 color matrix. The evaluations are carried out on a PDP11 computer. The factor scintigrams provide a quick method to look for abnormal regions in a nephrogram.

The evaluation of the perfusion studies is much more complex than the hippurate studies. The main problem seems to be the lack of an absolute time scale, i.e. the impossibility to assign t = 0 to a certain point. The activity arrives between 2 and more than 10s after starting of the data collection. These time differences influence the factor loadings considerably, without having any clinical meaning. To improve the results for this kind of study,

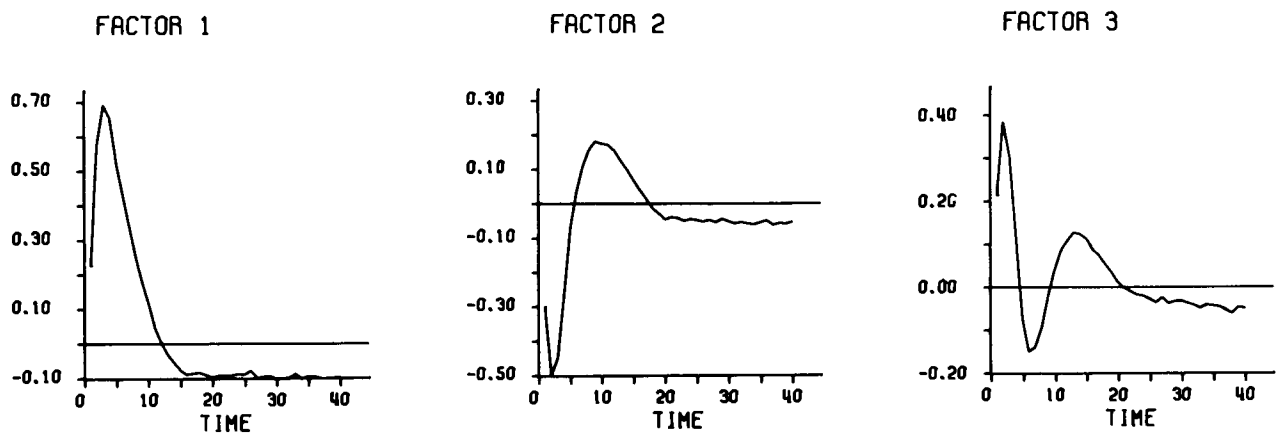


Fig.2: Plot of the three biggest factors for evaluation of kidney perfusion studies with ^{99m}-Tc-pertechnetate.

a check has been built in, which begins the evaluation at the point where the activity begins to come in. Yet the time interval between the incoming of the activity in the picture and in the kidney may still vary considerably.

Fig.3 shows the complexity of a pertechnetate-perfusion study. Mean factors were calculated over the kidney, sections of the arteries and the background. The motion of the bolus along the arteries shows up in the factors. A shift in the time scale would give similar effects.

The method was applied to different clinical problems. As an example, several transplanted kidneys were observed over periods up to three months. Here the factor evaluations of the hippurate-studies agree with the clinical findings concerning function and repulsion of the kidneys. The factor evaluations of the pertechnetate-studies give roughly correct results with some significant deviations.

Sharing of the procedure by different laboratories

Every type of study has its own reference factor set. Therefore laboratories using different versions of the same method may not necessarily use the same factor sets. The applicability of an existing factor set to a study can be tested by means of the fit of the reconstructed time-activity curves. If the factor set is sufficiently good, the sharing of the factors and of the programs is not difficult. It might be necessary for another laboratory which uses the same factor set to calibrate it by means of kidneys of well known properties. In any case every new set of factors has to be calibrated by means of a sufficient number of kidney studies.

Data compression

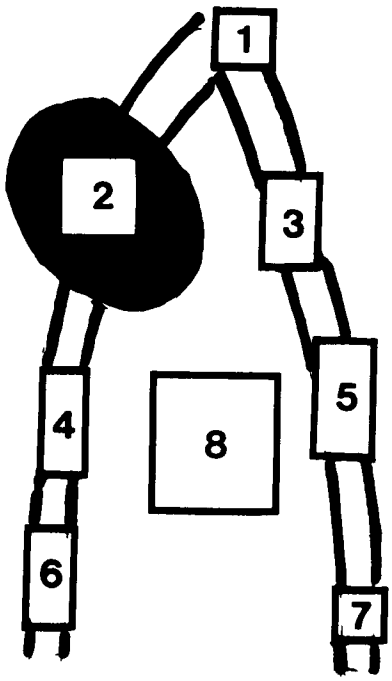
As a sequence scintigram can be reconstructed by means of the most important factors and factor loadings, the storage of factors and loadings together with the sum matrix n(x) instead of the complete sequence scintigram provides considerable saving of archival storage. A factor of 10 may easily be gained compared with sequential storage of data. There are two ways to perform factor storage: Either every study is used to calculate its own factor set. This takes more computing time, and the factor set has always to be stored together with the study, but every study is represented with sufficient accuracy. The other method uses the reference factor set to calculate the factor loadings. In this case the factors need not necessarily to be stored with every study. For studies with very abnormal time-activity functions it may be impossible to reconstruct the original data satisfactorily.

REFERENCES

1. P. Schmidlin, J. Clorius, R. Kubesch, K. Dreikorn, Evaluation of Dynamic Studies by Means of Factor Analysis, IAEA International Symposium on Medical Radionuclide Imaging, Los Angeles, Oct. 25-29, 1976.

2. J. P. Benzecri, L'analyse des Données (Dunod, Paris-Bruxelles-Montreal 1973).

3. S. Watanabe, Karhunen-Loève Expansion and Factor Analysis - Theoretical Remarks and Applications. In: Pattern Recognition (J. Sklansky editor, Dowden, Hutchinson & Ross Inc., Stroudsburg, Pennsylvania).



Patient 4560/76

Region	Factor 1	2	3
1	.227	-.117	.047
2	-.010	.101	.020
3	.052	-.023	-.042
4	.260	.040	-.164
5	.190	.073	-.174
6	.061	.125	-.131
7	.009	.175	-.066
8	-.219	-.111	-.017

Fig.3: Perfusion factor loadings of 8 regions of interest in an isotope nephrogram of a transplanted kidney.

DISCUSSION:

Line asked Dr. Schmidlin if he had compared the sensitivity of the factor loadings to kidney dysfunction, for example, in transplant rejection, with clinical parameters and with other functional maps.

Schmidlin responded by stating that factor analysis is, in principle, no different than the usual evaluation of the renogram curves; it is only very precise because the information from the whole curve is used to generate the factor loadings. He noted that he had discussed the problem of sensitivity with his clinician colleages but noted that there seems to be no single test which they used to determine rejection of a transplanted kidney. They always consider several clinical parameters to determine if a kidney is undergoing rejection. He indicated that he would like to put all of the usual clinical parameters and the three highest factor loadings into a vector and perform another factor analysis on this data to see which parameters are most important to describe rejection or improvement.

He indicated that they have already extracted the factors from 70 studies and have examined their relative magnitudes for various diseases and degrees of dysfunction. It is clear from this work that factor analysis will not be of much use for differential diagnosis but will be of value to concisely describe kidney function.

Anderson asked if it is possible to expand the factors in terms of orthogonal time functions so that they have more physical meaning.

Schmidlin did not think so and added that one may be able to say that one of the factors corresponds more to the kidney and another to the vasculature but he did not think it is possible to do better than this because the algorithm that he used to produce the factors is purely mathematical. One cannot expect that the factors produced by this algorithm will correspond to anything which has a direct anatomic interpretation or physical meaning.

IDENTIFICATION OF RENOGRAMS BY FACTOR ANALYSIS

B. E. Oppenheim
Department of Radiology, Indiana University School of Medicine
Indianapolis, Indiana, USA

ABSTRACT

Factor analysis was applied to a set of simulated renogram curves generated from a four-compartment model of the kidneys. Each curve could be reconstructed, with considerable accuracy, as the sum of the average renogram and a linear combination of a small number of orthogonal unit vectors. A curve is identified by its factors, which we define as the weights applied to the unit vectors in reconstructing the curve. For noisy renograms containing of the order of 1,000 counts the optimal number of factors is two. An efficient method is demonstrated for extracting parameters of shape from noisy renogram curves. This method is relatively insensitive to noise and appears well suited for the generation of functional images of the kidneys.

INTRODUCTION

Factor analysis is a powerful tool for observing interrelationships among variables and for obtaining economical descriptions of observed data. This technique has been used in statistics for many years to identify correlations between groups of variables[1]. Di Paola et al.[2] and Schmidlin and Rösel[3] have used this technique on nuclear images to obtain data-compression and image-smoothing. More recently, Schmidlin et al.[4] have applied factor analysis to clinically obtained renograms in order to separate normal from abnormal curves.

The present study was carried out to determine the value of factor analysis as a tool for identifying renogram shapes. In particular, it was felt that factor analysis should be useful in the generation of functional renal images. These images contain a parameter at each image point that characterizes the dynamic events occurring at that point[5]. The dynamic events for a renal study are represented by time-activity curves (renograms), and a separate curve can be generated for each image point by acquiring a dynamic study with a scintillation camera, storing the data digitally in list or frame mode, and reformatting the data. Some parameter of shape, such as time-to-peak, is evaluated for each curve, and that value is assigned to the corresponding image point to form the functional image. Each curve is very noisy, however, since it is formed from events collected in a very small portion of the crystal of the scintillation camera. One would like to apply noise filtering or some curve-fitting scheme, but is limited by the large number of curves that must be processed

(about 4,000 for a 64 x 64 image). Factor analysis provides a rapid technique for curve fitting, and one goal of this study was to develop an efficient method for estimating parameters from noisy renogram curves, using this technique.

The main objective of this study was to determine the accuracy with which noisy renograms could be identified by factor analysis. It was first necessary to establish a data base of renograms. This was done by generating simulated renograms with the computer, based on a four-compartment model. Factor analysis was applied to these renograms to generate a set of basis vectors, which could be used to compute the factors corresponding to any given renogram. A renogram can be reconstructed from its factors, but the reconstruction is imperfect, especially if the renogram is noisy, so an analysis of error was carried out to determine the accuracy of the reconstruction in the presence of noise. Finally, a method employing factor analysis was compared with several other methods for estimating the time-to-peak parameter of noisy renograms.

RENOGRAM DATA BASE

Simulated renograms were used for this study since the analysis of error required knowledge of the true shape of the renogram. These were generated from a four-compartment model, which is a simplification of the seven-compartment model proposed by DeGrazia et al.[6]. This model consists of a plasma pool, a single renal mixing pool, a single renal delay pool, and a bladder pool (Fig. 1). It does not include an extravascular pool or the second renal mixing and delay pools.

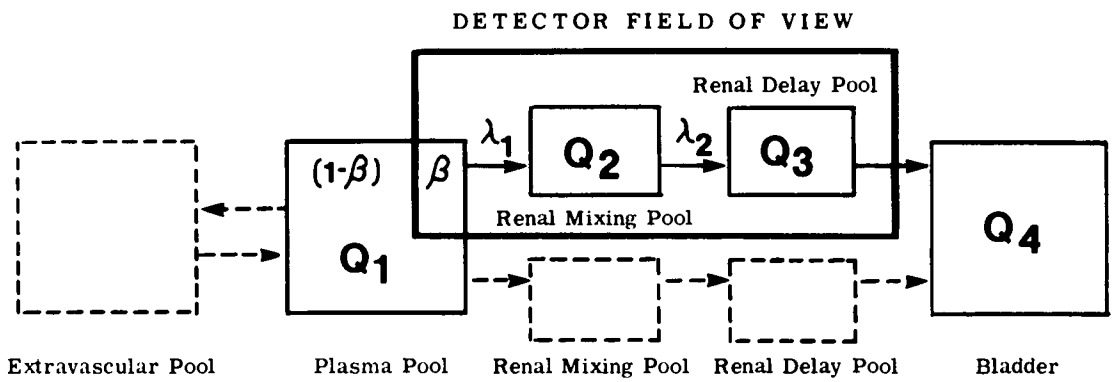


Fig. 1. Four-compartment model used in this study. Dashed rectangles are three additional compartments included in the model of DeGrazia et al.[6] The field of view of the detector determines the portion of the activity which is detected to produce the renogram.

The equations for the model are:

Plasma pool: $$\dot{Q}_1(t) = -\lambda_1 Q_1(t) \tag{1}$$

Renal mixing pool: $$\dot{Q}_2(t) = \lambda_1 Q_1(t) - \lambda_2 Q_2(t) \tag{2}$$

Renal delay pool: $$\begin{aligned} \dot{Q}_3(t) &= \lambda_2 Q_2(t) - \lambda_2 Q_2(t-t_d) && \text{for } t \geq t_d \\ &= \lambda_2 Q_2(t) && \text{for } t \leq t_d \end{aligned} \tag{3}$$

Bladder: $$\begin{aligned} \dot{Q}_4(t) &= \lambda_2 Q_2(t-t_d) && \text{for } t \geq t_d \\ &= 0 && \text{for } t \leq t_d \end{aligned} \tag{4}$$

with initial conditions

$$Q_1(0) = I \qquad Q_2(0) = 0 \qquad Q_3(0) = 0 \qquad Q_4(t \leq t_d) = 0$$

where Q_i is the quantity of tracer in pool i, $\dot{Q}_i$ is its rate of change, λ_1 and λ_2 are the rate constants for clearance from the plasma and renal mixing pools respectively, t_d is the delay time constant introduced by the renal delay pool, and I is the injected dose. In some formulations[6-9] the mixing pool precedes the delay pool; in others[10-11] it follows it. It can be shown that for the four-compartment model identical results are obtained in both cases.

The solutions to these equations are:

$$Q_1(t) = Ie^{-\lambda_1 t} \tag{5}$$

$$Q_2(t) = I[\lambda_1/(\lambda_2-\lambda_1)](e^{-\lambda_1 t} - e^{-\lambda_2 t}) \tag{6}$$

$$Q_3(t) = P(t) - P(t-t_d) \tag{7}$$

$$Q_4(t) = P(t-t_d) + I \tag{8}$$

where $$\begin{aligned} P(t) &= I(\lambda_1 e^{-\lambda_2 t} - \lambda_2 e^{-\lambda_1 t})/(\lambda_2-\lambda_1) && \text{for } t \geq t_d \\ &= -I && \text{for } t \leq t_d \end{aligned}$$

The equation for the renogram is given by

$$r(t) = K[\beta Q_1(t) + Q_2(t) + Q_3(t)] \tag{9}$$

where β is the fraction of the plasma pool within the detector's field of view and K is the gain of the detector (see Fig. 1). Substituting Eqs. 5-7 in Eq. 9, we have

$$\begin{aligned} r(t) &= KI\left[\frac{t_p e^{-(t-t_d)/t_p} - t_m e^{-(t-t_d)/t_m}}{t_p - t_m} - (1-\beta)e^{-t/t_p}\right] && \text{for } t \geq t_d \\ &= KI\left[1 - (1-\beta)e^{-t/t_p}\right] && \text{for } t \leq t_d \end{aligned} \tag{10}$$

where $t_p = 1/\lambda_1$ is the plasma clearance time constant and $t_m = 1/\lambda_2$ is the mixing pool time constant. The terms K and I affect the magnitude of the renogram but not its shape. We will be using factor analysis only to identify shape,

as magnitude information is readily available from summed images of a dynamic study, and we may therefore ignore K and I. We thus have a general expression for the renogram in terms of the four physiological parameters t_p, t_m, t_d and β.

A set of 512 curves was generated by Eq. 10 using all possible combinations of the four parameters, where these parameters were allowed to take on the following values:

$$\begin{aligned} t_p &= 2^{A-2} \text{ min} & A &= 2,3,4,5,6,7,8,9 \\ t_m &= 2^{B-2} \text{ min} & B &= 0,1,2,3,4,5,6,7 \\ t_d &= 2^{C-2} \text{ min} & C &= 2,3,4,5 \\ \beta &= .1\,D & D &= 1,2 \end{aligned} \tag{11}$$

This set was observed to span the entire range of renograms likely to be encountered clinically. Each curve was assigned a four-digit code ABCD, where the digits identify the parameters according to Eq. 11 (e.g., curve 6232 has parameters t_p = 16 min., t_m = 1 min., t_d = 2 min. and β = .2).

Background time-activity curves will be obtained when imaging non-renal portions of the abdomen. The corresponding curves are derived from Eq. 5:

$$r(t) = K\beta Q_1(t) = K\beta I e^{-\lambda_1 t} = K\beta I e^{-t/t_p} \tag{12}$$

Since we are again only interested in curve shape we can ignore K, β and I. The background curves are therefore functions of the single parameter t_p. Eight background curves were generated by Eq. 12, and each was assigned a code A000, where integer A identifies the value of t_p as in Eq. 11.

Upon visual inspection of the 512 renogram curves, 162 were selected as curves that might occur clinically. Of the eight background curves, only the five for which $A \geq 5$ were felt to be likely to occur clinically. The renogram data base should be somewhat representative of the relative frequencies with which the various curves will occur, and since in functional imaging the majority of curves encountered will be background curves, it was decided to include each background curve A000 as many times as there were renogram curves with first term A, with the restriction that $A \geq 5$. This resulted in 137 background curves, which together with the 162 renogram curves form the data base for this study. Each curve in the data base was normalized to unit area by dividing each of its terms by the sum of the terms. Examples of eight of these curves are given in Fig. 2.

FACTOR ANALYSIS

We are given m renogram curves in the data base (henceforth the word "renogram" will apply to both renogram and background curves, unless otherwise indicated). We represent each by an n-dimensional vector $\vec{r}_j = [r_{ij}]$ (i=1,..,n), where n is the number of time intervals during which the renogram is evaluated, and the r_{ij} are the values for each interval. Since each renogram has been normalized to unit area we have $\Sigma_i r_{ij} = 1$. We define $\vec{r}_{av} = \frac{1}{m} \Sigma_j \vec{r}_j$ as the average renogram. We form the set of shifted renograms $\vec{r}_j' = \vec{r}_j - \vec{r}_{av}$ (j=1,..,m). Next, the autocorrelation matrix $C = [c_{ik}]$ (i,k=1,..,n) is formed, where $c_{ik} = \Sigma_j r_{ij}' r_{kj}'$. We compute the orthogonal eigenvectors $\vec{e}_\ell$ (ℓ=1,..,n) of C, normalized to be unit vectors and ordered by decreasing magnitude of the corres-

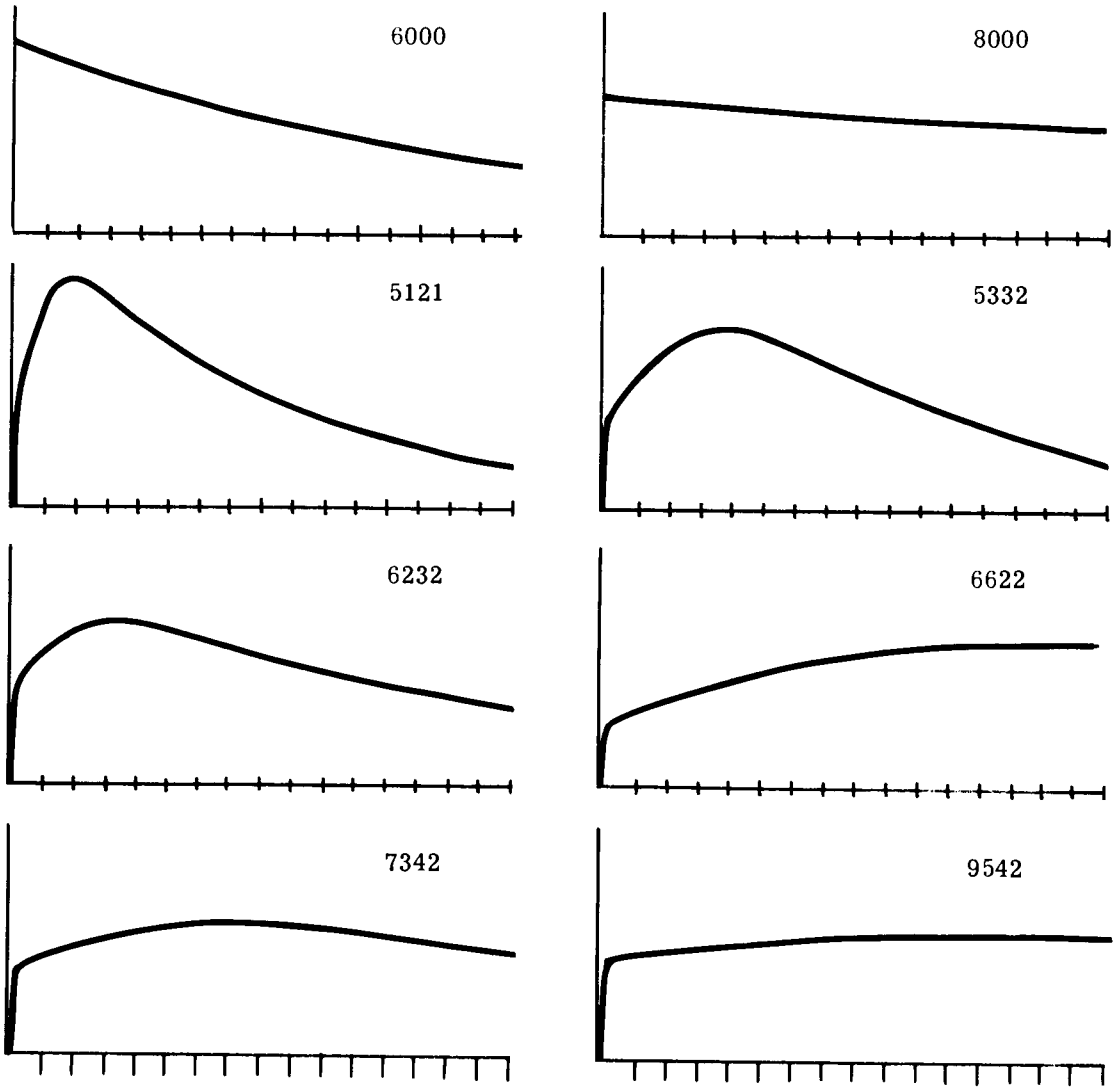


Fig. 2. Two typical background curves and six typical renogram curves generated by the four-compartment model, with their four-digit codes.

ponding eigenvalues. These vectors form a basis for n-dimensional space. Then any normalized renogram $\vec{r}$ has factors* f_ℓ (ℓ=1,..,n), where

$$f_\ell = (\vec{r} - \vec{r}_{av})^T \cdot \vec{e}_\ell \tag{13}$$

Furthermore, $\vec{r}$ can be approximated by $\vec{r}^*_p$, the reconstruction using the first p factors:

$$\vec{r}^*_p = \vec{r}_{av} + \sum_{\ell=1}^{p} f_\ell \vec{e}_\ell \tag{14}$$

The approximation improves as p increases and becomes an identity for p=n.

As an alternative, the autocorrelation matrix, the eigenvectors, and the factors could have been generated from the unshifted renograms $\vec{r}$. In this case the equation for $\vec{r}^*_p$ would have been given by

$$\vec{r}^*_p = \sum_{\ell=1}^{p} f_\ell \vec{e}_\ell \tag{15}$$

It is found, however, that the term $f_1\vec{e}_1$ in Eq. 15 adds only about as much accuracy to the reconstruction of $\vec{r}$ as the term $\vec{r}_{av}$ in Eq. 14, so that one needs to use one less factor with Eq. 14 than with Eq. 15 to obtain comparable accuracy. It is desirable to reduce the number of factors to achieve computational efficiency. Also, statistical noise in the reconstructed renogram increases with the number of factors, as shown below. The approach employing the shifted renograms (Eq. 14) is therefore preferred.

Since the curves in the data base contain little high frequency information they can be coarsely sampled with little loss of information. Each curve was therefore subdivided into sixteen one minute intervals and was represented by a vector $\vec{r} = [r_i]$ (i=1,..,16), where r_i represents the area under the curve for the i-th interval (as determined from integrals of Eqs. 10 or 12 with respect to t) and $\Sigma r_i = 1$. From these vectors the average vector $\vec{r}_{av}$ was determined and the other steps described above were carried out, yielding sixteen 16-dimensional unit eigenvectors $\vec{e}_\ell$. Figure 3 shows $\vec{r}_{av}$ and the first three eigenvectors.

The sixteen factors f_ℓ were computed for each renogram in the data base. It was generally found that the higher the factor number, the smaller the magnitude of the factor. For each factor number ℓ the mean value of the magnitude of f_ℓ for 162 renograms, $\overline{|f_\ell|}$, was determined. This value is plotted as a function of ℓ in Fig. 4.

ERROR IN FACTORS

Let $\vec{r} = [r_i]$ be a renogram generated by Eq. 10 and normalized so that $\Sigma r_i = 1$. Its ℓ-th factor f_ℓ is given by Eq. 13. We first wish to determine the error in f_ℓ when noise is introduced into $\vec{r}$.

For some large integer N let $\vec{R} = [R_i] = [Nr_i]$ be the N-count noiseless renogram corresponding to $\vec{r}$. We introduce noise by defining the Poisson stochastic process $\vec{\mathbb{R}} = [\mathbb{R}_i]$, where $E\{\mathbb{R}_i\} = Var\{\mathbb{R}_i\} = R_i$. The notation $\mathbb{X}$ indicates

*The terminology of factor analysis is somewhat confusing. We will refer to the vectors as "eigenvectors" and their weights as "factors". Harman[1] calls them "factors" and "factor loadings", respectively, while Di Paola[2] and Schmidlin[4] refer to them as "G-factors" and "F-factors".

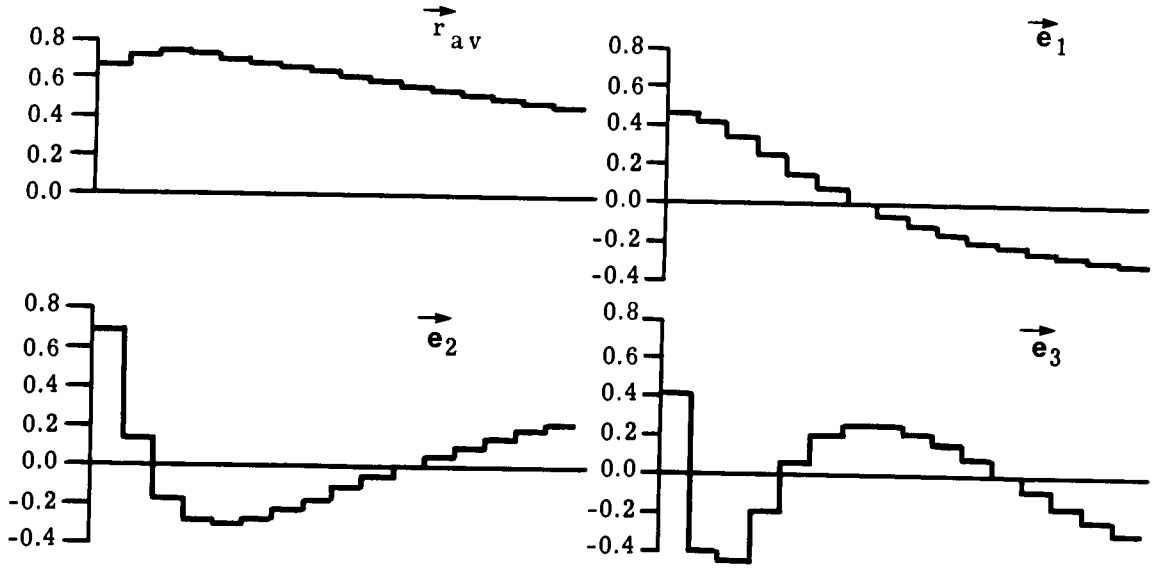


Fig. 3. Average renogram and first three eigenvectors for 299 curves in renogram data base.

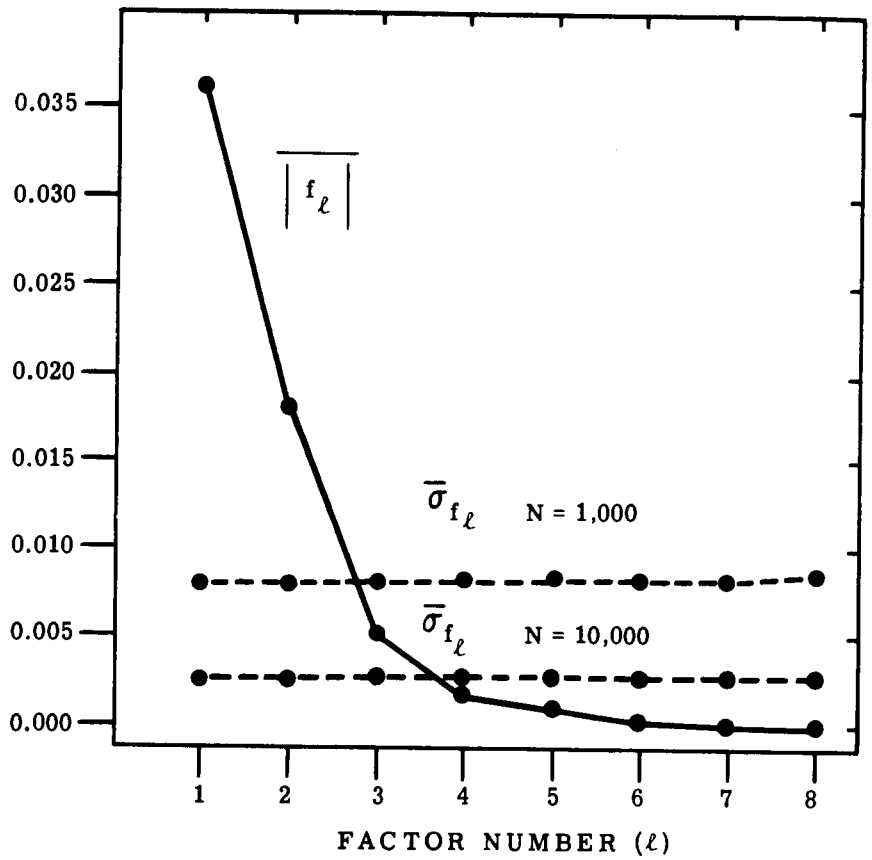


Fig. 4. Mean value of magnitude of ℓ-th factor (solid line) and mean value of standard error of ℓ-th factor (dashed lines) over 162 renogram curves, for first eight factors. Value of N is total counts in renogram.

that X is a random variable, E{ } is the expected value operator, and Var{ } is the variance operator. We normalize $\vec{\underset{\sim}{R}}$ by letting $\vec{\underset{\sim}{r}} = [\underset{\sim}{r}_i] = [\underset{\sim}{R}_i/\underset{\sim}{N}]$ where $\underset{\sim}{N} = \Sigma \underset{\sim}{R}_i$. Then

$$E\{\underset{\sim}{r}_i\} = E\{\underset{\sim}{R}_i\}/E\{\underset{\sim}{N}\} = R_i/N = r_i$$

Now since $\mathrm{Var}\{\underset{\sim}{x}/\underset{\sim}{y}\} \simeq (\mu_x/\mu_y)^2[(\sigma_x/\mu_x)^2 + (\sigma_y/\mu_y)^2]$ provided that $\sigma_x \ll \mu_x$ and $\sigma_y \ll \mu_y$, we have

$$\mathrm{Var}\{\underset{\sim}{r}_i\} \simeq r_i^2(1/R_i + 1/N) = r_i(1 + r_i)/N \tag{16}$$

Now the ℓ-th factor $\underset{\sim}{f}_\ell$ of $\vec{\underset{\sim}{r}}$ is given by

$$\underset{\sim}{f}_\ell = (\vec{\underset{\sim}{r}} - \vec{r}_{av})^T \cdot \vec{e}_\ell \tag{17}$$

Then
$$E\{\underset{\sim}{f}_\ell\} = (E\{\vec{\underset{\sim}{r}}\} - \vec{r}_{av})^T \cdot \vec{e}_\ell = (\vec{r} - \vec{r}_{av})^T \cdot \vec{e}_\ell = f_\ell \tag{18}$$

And
$$\mathrm{Var}\{\underset{\sim}{f}_\ell\} = \mathrm{Var}\{\vec{\underset{\sim}{r}}^T \vec{e}_\ell\} = \mathrm{Var}\{\Sigma_i \underset{\sim}{r}_i e_{i\ell}\} = \Sigma_i e_{i\ell}^2 \mathrm{Var}\{\underset{\sim}{r}_i\}$$

since $\underset{\sim}{r}_i$ and $\underset{\sim}{r}_j$ may be considered uncorrelated for $i \neq j$ (they are very weakly correlated through $\underset{\sim}{N}$). Applying Eq. 16,

$$\mathrm{Var}\{\underset{\sim}{f}_\ell\} \simeq \Sigma_i e_{i\ell}^2 r_i(1 + r_i)/N \tag{19}$$

Then σ_{f_ℓ}, the standard error of f_ℓ, is the square root of this.

The value of σ_{f_ℓ} was computed for each renogram for each factor number ℓ, for several values of N. It was found that for a given N there was little variability in σ_{f_ℓ} for different ℓ or for different renograms. The average value of σ_{f_ℓ} over 162 renograms, designated by $\overline{\sigma_{f_\ell}}$, is plotted in Fig. 4 for each ℓ and for N=1,000 and N=10,000. Note than when N=1,000 the value of $\overline{\sigma_{f_\ell}}$ exceeds the average magnitude of f_ℓ for factors beyond the second, and when N=10,000 this occurs for factors beyond the third. One would expect that these higher numbered factors would add more error than information in the reconstruction of renograms. This proves to be the case, as is shown below.

ERROR IN RECONSTRUCTION OF RENOGRAMS

For the noiseless normalized renogram $\vec{r}$ the reconstruction from p factors, $\vec{r}_p^*$, is given by Eq. 14. For the corresponding noisy normalized renogram $\vec{\underset{\sim}{r}}$ generated from a renogram with N expected counts, the reconstruction from p factors is given by:

$$\vec{\underset{\sim}{r}}_p^* = \vec{r}_{av} + \sum_{\ell=1}^{p} \underset{\sim}{f}_\ell \vec{e}_\ell \tag{20}$$

We are interested in determining how well $\vec{r}$ is approximated by $\vec{\underset{\sim}{r}}_p^*$. We define the error in $\vec{\underset{\sim}{r}}_p^*$ as

$$\underset{\sim}{\delta}_p = |\vec{\underset{\sim}{r}}_p^* - \vec{r}|/|\vec{r}| \tag{21}$$

and let $\delta_p = E\{\underset{\sim}{\delta}_p\}$. We wish to determine how δ_p behaves as a function of p and N. From Eq. 21,

$$\begin{aligned} E\{\underset{\sim}{\delta}_p^2\} &= E\{|\vec{\underset{\sim}{r}}_p^* - \vec{r}|^2\}/|\vec{r}|^2 = E\{|(\vec{\underset{\sim}{r}}_p^* - \vec{r}_p^*) + (\vec{r}_p^* - \vec{r})|^2\}/|\vec{r}|^2 \\ &= E\{|\vec{\underset{\sim}{r}}_p^* - \vec{r}_p^*|^2\}/|r|^2 + |\vec{r}_p^* - \vec{r}|^2/|\vec{r}|^2 \end{aligned} \tag{22}$$

since $E\{\underset{\sim}{\vec{r}}_p^*\} = \vec{r}_p^*$, as follows from Eqs. 14, 18 and 20. From these same equations we have

$$E\{|\underset{\sim}{\vec{r}}_p^* - \vec{r}_p^*|^2\} = E\{\sum_{\ell=1}^{p} (\underset{\sim}{f}_\ell - f_\ell)^2\} = \sum_{\ell=1}^{p} \mathrm{Var}\{\underset{\sim}{f}_\ell\} \tag{23}$$

Using Eqs. 14, 19, 22 and 23 we have

$$E\{\underset{\sim}{\delta}_p^2\} \simeq \frac{1}{|\vec{r}|^2}\left[\frac{1}{N}\sum_{\ell=1}^{p}\sum_{i=1}^{n} e_{i\ell}^2 r_i(1+r_i) + \left|\sum_{\ell=1}^{p} f_\ell \vec{e}_\ell - (\vec{r}-\vec{r}_{av})\right|^2\right] \tag{24}$$

Then $\delta_p = E\{\underset{\sim}{\delta}_p\}$ is approximated by the square root of Eq. 24. This is reasonable, for even when δ_p is small $\underset{\sim}{\delta}_p$ must always be positive or zero by Eq. 21.

It is seen from Eqs. 22 and 24 that there are two components to δ_p, the expected error in $\underset{\sim}{\vec{r}}_p^*$. The first term represents the error due to statistical noise in $\underset{\sim}{\vec{r}}_p^*$, and decreases as N, the number of counts used to produce $\underset{\sim}{\vec{r}}$, increases. The term increases, however, as the number of factors (=p) increases. The second term represents the accuracy of the noiseless reconstruction $\vec{r}_p^*$, and decreases as p increases.

The value of δ_p was calculated for each renogram for various combinations of p and N, and the values for each combination were averaged over 162 renograms to give $\overline{\delta_p}$. The results are shown in Fig. 5. As expected, for N=1,000 the most accurate reconstructions are obtained when two factors are used, while for N=10,000 a slight improvement is obtained by adding the third factor.

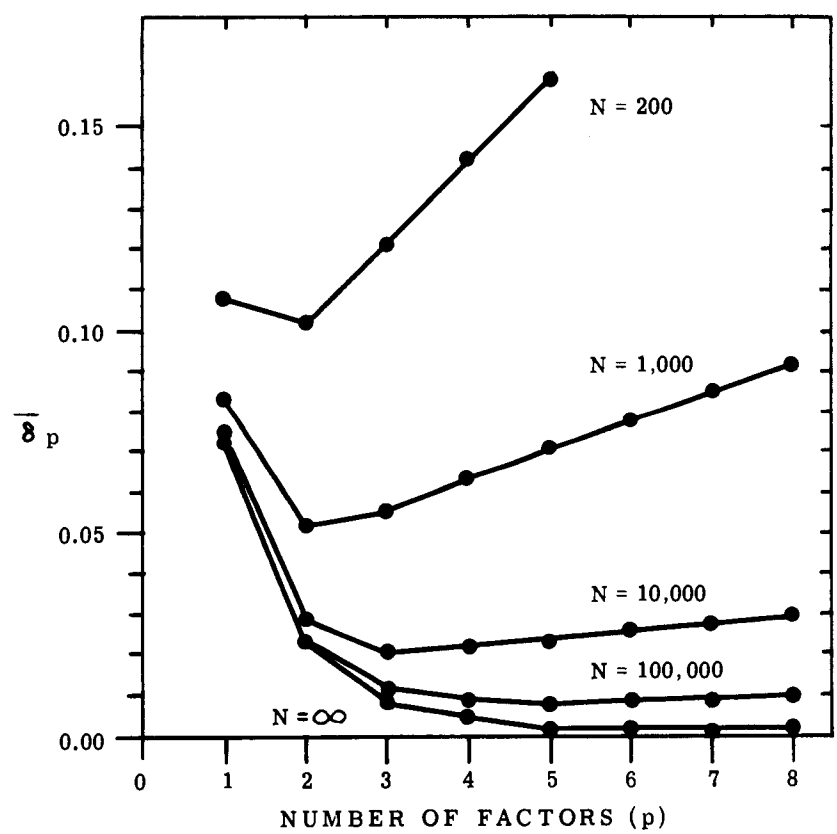


Fig. 5. Mean value over 162 renograms of expected error in reconstruction of noisy renogram, as function of number of factors used for reconstruction (p) and total number of counts in renogram (N).

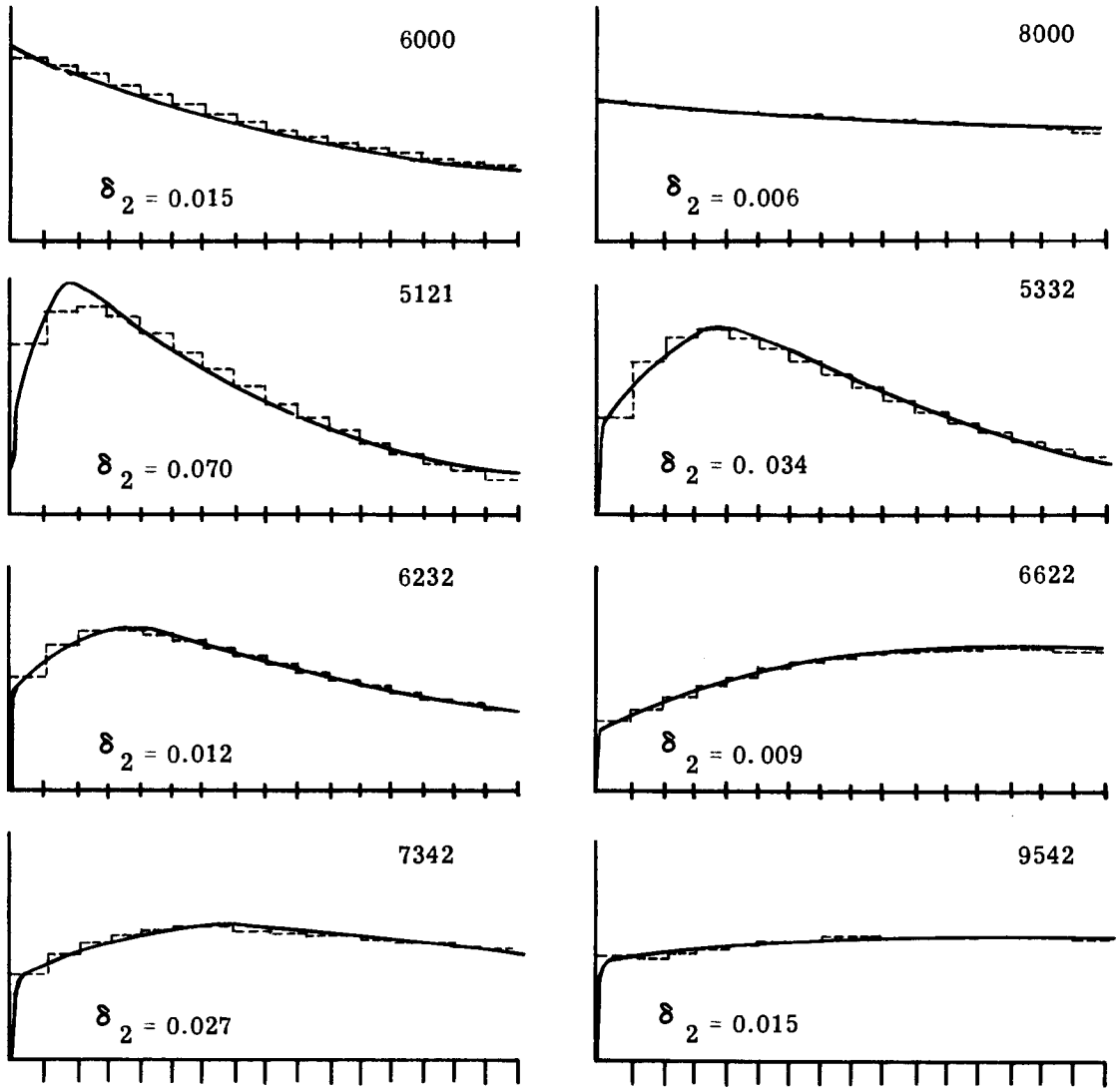


Fig. 6. Reconstructions of curves in Fig. 2 from two factors, with values of error in reconstruction.

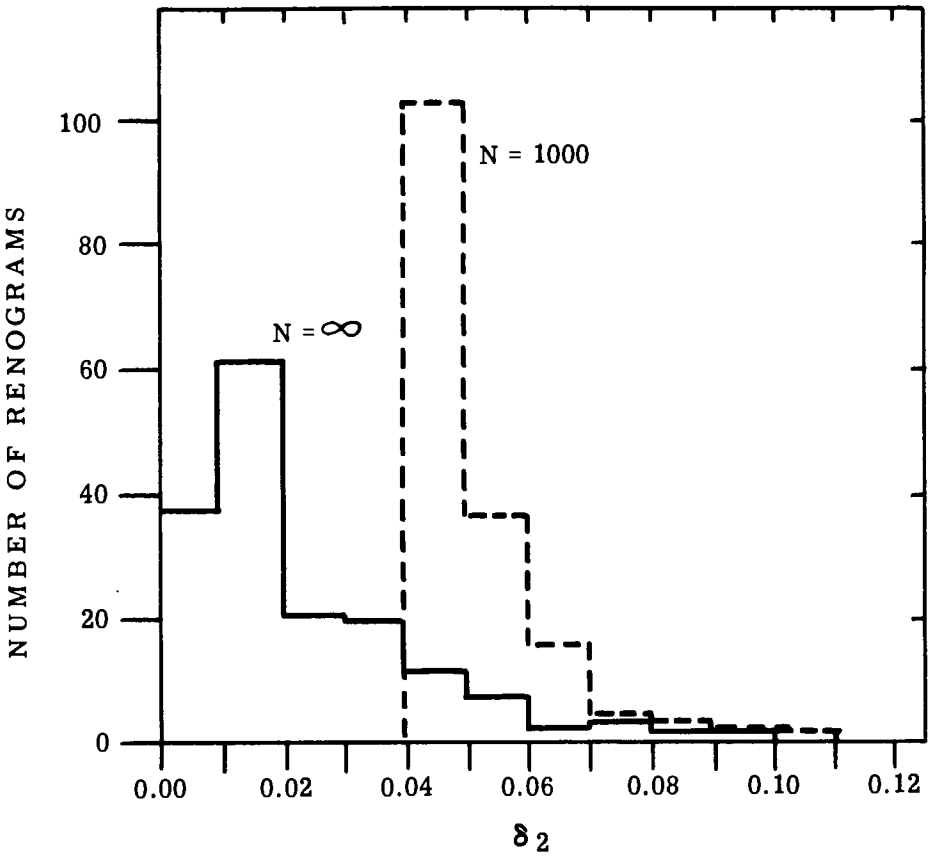


Fig. 7. Distribution (among 162 renograms) of expected error in reconstruction from two factors, for noiseless renograms and renograms containing 1,000 total counts.

It is anticipated that the renograms that will be used to generate functional images will contain of the order of 1,000 total counts, and therefore two factors should be appropriate. Reconstruction of noiseless renograms from two factors are demonstrated in Fig. 6, along with the values of δ_2, the errors in the reconstructed renograms, which are seen to be generally quite small. Figure 7 shows the distribution of the expected error among the 162 renograms for infinite counts and for N=1,000. The presence of noise increases the error; nevertheless for N=1,000 the value of δ_2 is nearly always between 0.04 and 0.07, which represents a reasonably good reconstruction.

ESTIMATION OF CURVE PARAMETERS

Since factor analysis enables one to identify curve shapes it can be used to estimate parameters of curve shape. As an example, a method based on factor analysis was devised for determining t_{max}, the time-to-peak parameter of a renogram, when the renogram is very noisy. This method was implemented as follows:

The ranges of the first two factors were determined from the renogram data base. These were -0.08 to +0.10 for the first factor and -0.06 to +0.02 for the second factor. Both ranges were sampled with a sampling interval of 0.25, giving 72 values for the first factor and 32 for the second. Note that although the range of f_1 was more than twice the range of f_2, the same sampling interval was employed for both, since the standard errors of f_1 and f_2 are about the same, as shown in Fig. 4.

For each pair of factor values the corresponding 16-interval renogram was generated by Eq. 14. A cubic was fit to the renogram in the region of the maximum, using a modified cubic spline technique[12] and taking into account the fact that the renogram values are interval rather than point values (a parabola was fit when the maximum was the first, next to last or last term of the renogram). Then t_{max} for the pair of factors was the x-coordinate of the maximum of the cubic, with the restriction that $0 \leq t_{max} \leq 16$. This value was entered into a 72 x 32 table of t_{max} as a function of the values of the first two factors, which we will call the factor table. This table can be used to estimate t_{max} for a given noisy renogram in the following manner: The first two factors for the renogram are determined by Eq. 13. The value of t_{max} is then found from the factor table, using 2-dimensional linear interpolation.

An experiment was carried out to compare the accuracy of the estimate of t_{max} obtained from the factor table with the accuracy of the estimates obtained by direct inspection of the noisy curve and a smooth version of that curve. This experiment was applied to each curve in the renogram data base in the following manner:

For a given renogram ABCD the true value of t_{max} (found analytically from Eq. 10) is:

$$t_{max} = \frac{t_p t_m}{t_m - t_p} \left[\ln\{e^{t_d/t_p} - (1-\beta)(1-\frac{t_m}{t_p})\} - \frac{t_d}{t_m} \right] \tag{25}$$

where the parameters are evaluated according to Eq. 11. For each background curve A000 the true value of t_{max} is 0. The corresponding 64-interval curve vector $\vec{r}$ was generated from an integrated form of Eq. 10 or 12 and normalized so that the sum of its values was unity. One hundred random renograms $\vec{\underset{\sim}{R}}$ were generated for $\vec{r}$, with N=1,000, by letting $\vec{R} = [R_i] = [Nr_i]$ (i=1,..,64) and replacing R_i by random numbers $\underset{\sim}{R}_i$ such that $E\{\underset{\sim}{R}_i\} = Var\{\underset{\sim}{R}_i\} = R_i$, using a Gaussian random number generator.

For each $\vec{\underset{\sim}{R}}$ the value of t_{max} was found by direct inspection. A smoothed renogram was then generated by applying a seven point smoothing algorithm (based on the binomial distribution) to $\vec{\underset{\sim}{R}}$, and its t_{max} was found by direct inspection. Then $\vec{\underset{\sim}{R}}$ was compressed to a 16-interval vector and normalized by dividing each value by the total counts in $\vec{\underset{\sim}{R}}$, factors f_1 and f_2 were found for the normalized vector by Eq. 13, and the estimated value of t_{max} was obtained from the factor table by 2-dimensional linear interpolation.

Figure 8 shows the mean and one standard deviation range of the estimates of t_{max} obtained by each of these approaches together with the true value of t_{max}, for eight selected curves, where N=1,000. The entire process was repeated for N=10,000and these results are shown in Fig. 9. The mean estimate of t_{max} from the factor table was, on the whole, little better than the mean estimate obtained by inspection of the unsmoothed and smoothed curves. The factor table method was, however, much more effective than the other methods in reducing the variability of t_{max}.

The results shown in Figs. 8 and 9 are representative of the results for all 162 renogram curves and the five background curves. For N=1,000 the mean estimate of t_{max} by the factor table method was better than the mean estimates by the other methods about two-thirds of the time. For N=10,000 it was better about one-third of the time. The standard error was smaller for the factor table method in nearly every instance, and significantly smaller in most instances. In one situation, however, the factor table method occasionally gave misleading results. When the true renogram was fairly flat with an early peak the corresponding noisy renogram would sometimes be mistaken for a background curve.

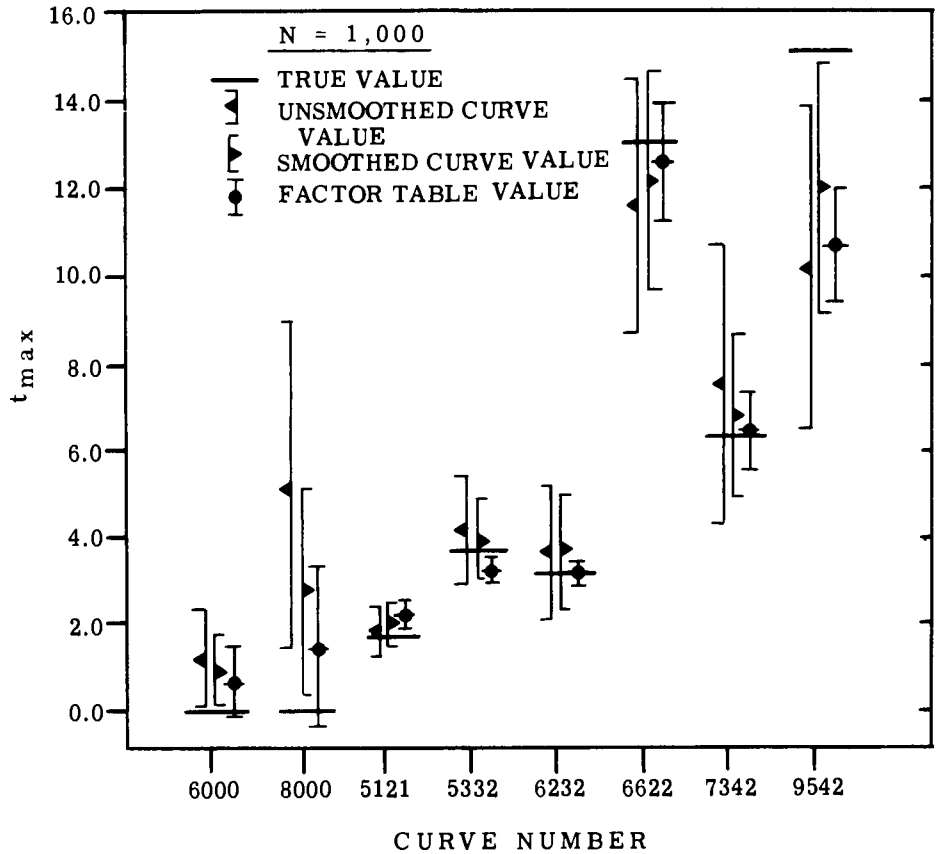


Fig. 8. True value of t_{max} and estimates for noisy curves obtained by three methods: direct inspection of unsmoothed curve, inspection of curve smoothed by weighted seven point smoothing, and factor table method. One hundred random curves, each with 1,000 expected total counts, were generated for each curve in Fig. 2 and mean value and one standard deviation range of estimates of t_{max} were determined for each method.

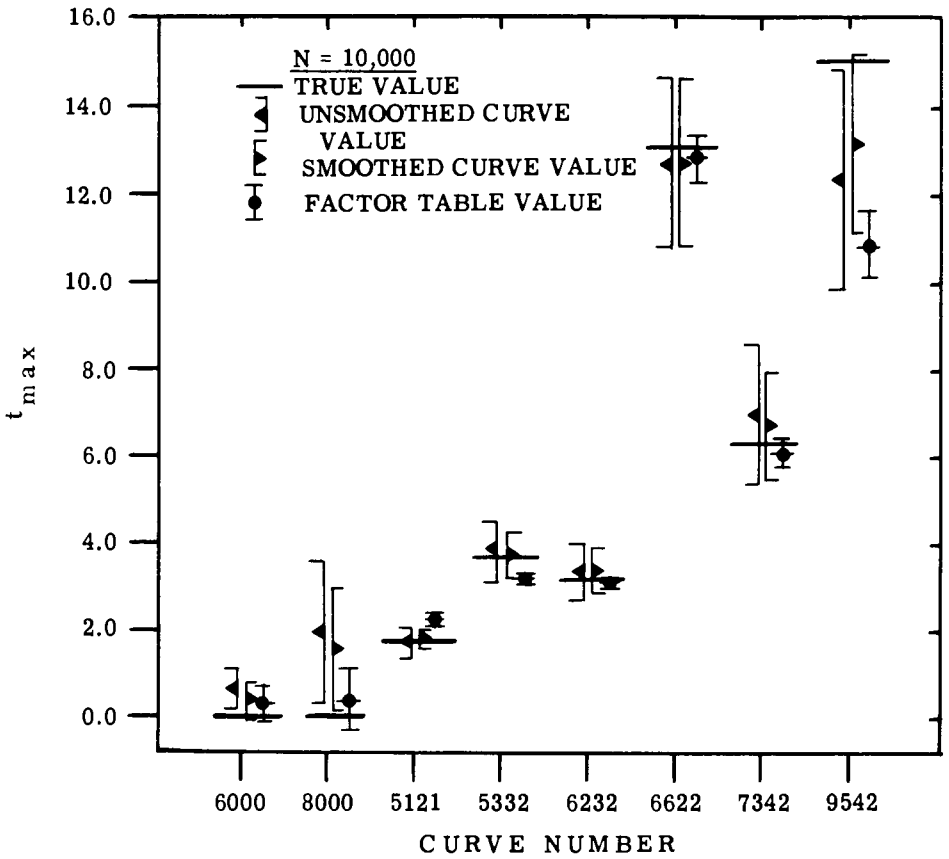


Fig. 9. Same as Fig. 8, except each random curve had 10,000 expected total counts.

With regard to computational efficiency, the most efficient method was, of course, direct inspection of the unsmoothed curve, as it required only comparisons. The factor table method was much more efficient than the method of smoothing followed by inspection, as it required only 54 multiplications to identify t_{max} (16 to normalize the renogram, 32 to compute and two to scale the two factors, and four for 2-dimensional interpolation), while seven-point smoothing of a 64 interval renogram required 262 multiplications (256 for smoothing plus six to adjust each of the outer three values). The cost of generating the factor table may be ignored in this comparison since this need only be done once.

DISCUSSION

Factor analysis provides an economical means for describing the observed values of highly correlated variables. We can consider a renogram evaluated during n time intervals to represent one set of outcomes of n variables. In general, given a collection of n-dimensional vectors, each representing a set of outcomes of n variables, one extracts the eigenvectors of the autocorrelation matrix, which form an ordered set of orthogonal basis vectors that span n-dimensional space. According to Metz[13] this basis set is optimal in the sense that for any $p \leq n$, of all possible p-dimensional subspaces the one spanned by the first p basis vectors is the one that best fits (in the least squares sense) the original collection of vectors. Now when the n variables are highly correlated the original vectors will be nearly completely included in some p-dimensional subspace of much lower dimension than n. Then each of these vectors can be nearly exactly reconstructed as a linear combination of the first p basis vectors, and therefore can be described by the p numbers (the so-called factors) which are the weights applied to the basis vectors. Thus the number of numbers required to describe each vector is reduced from n to p.

In the present study each of the renograms can be described rather well by only two factors, since it can be reconstructed fairly accurately from these two factors by Eq. 14, as is illustrated in Fig. 6. Moreover, two does appear to be the optimal number of factors for approximating noisy renograms where the total number of counts is of the order of 1,000. This is fortunate, for computation time increases linearly and memory required for storing a factor table increases nearly exponentially with the number of factors.

Factor analysis can be used for cataloging renograms. Schmidlin et al.[4] have shown that certain ranges of factor values are associated with pathology. It may be possible to refine this approach so that certain combinations of factor values imply specific types of abnormalities (e.g., obstruction or impaired renal function).

The factor table method described here offers an efficient means for estimating parameters of renogram shape in the presence of noise. This method should be useful for generating functional images. The parameter t_{max}, the time to maximum value, has been found to be a useful indicator of unilateral renal artery stenosis[14], and has been used for functional imaging[5,15]. The present study has shown that the estimates of t_{max} by the factor table method are as accurate as those obtained by direct inspection of the unsmoothed or smoothed renograms, but are much less sensitive to noise. Functional images of t_{max} based on these estimates should therefore be as accurate but less noisy than functional images obtained by other methods. A likely explanation for the noise reduction obtained by the factor table method is that this method fits the entire curve, while the other methods concentrate on a small portion of curve and are therefore more subject to noise. It is anticipated that other parameters of renogram shape, such as those reflecting washout from the kidneys, will similarly demonstrate the advantage of the factor table method.

Another potential application of factor analysis is in the identification of different functional components of the dynamic renal image (i.e., background, renal cortex, renal pelvis, bladder). This is currently being done by the tedious process of inspecting the images and manually flagging these components. The renograms generated from each component will often contain more information than the total renogram[16-19]. Halko et al.[20] have elegantly demonstrated that factor analysis with varimax rotation can identify discrete functional components within each kidney. One would expect their method to be very slow, especially if the number of renograms to be analyzed is large (see Harman[1], p. 291). Wiener et al.[21] have demonstrated a more practical approach, in which the various functional components are identified by determining the rate of change of radioactivity concentration in each region of the image during selected time intervals. This method requires least squares fitting of lines to segments of the renogram curves. Factor analysis should accomplish this end more rapidly and accurately by associating renogram shapes with functional components. It will probably be desirable to use magnitude information (obtained directly from a summed image) together with shape information (obtained through factor analysis) in identifying the functional components of the image.

Several remarks about the renogram model are in order. A four-compartment model was used because it permitted the generation of renograms from only four parameters of renal function, and because it permitted the explicit solution of the renogram equation (Eq. 10) and the equation for t_{max} (Eq. 25). The clinical situation is more accurately described by the seven-compartment model of DeGrazia et al.[6]. This model contains three additional compartments, which are indicated by dashed lines in Fig. 1. If the extravascular pool had been included in the present model, the peak of each curve would have been slightly lower and the terminal portion of each curve would have been slightly higher. The second renal mixing and renal delay pools would have had no effect on renogram shape if the two kidneys functioned equally or one did not function at all; otherwise they would have modified the shapes of the curves slightly. Because these differences are small the curves generated by the four-compartment model are sufficiently similar to real renogram curves, especially in the presence of noise, that the model suffices for the theoretical study carried out here. When applying factor analysis to clinical data, however, it would probably be desirable to construct the renogram data base from real curves as Schmidlin et al.[4] have done, or from curves generated by the more accurate model of DeGrazia et al.[6].

CONCLUSIONS

Factor analysis can be used to identify renogram shapes and extract curve parameters in the presence of noise. When total counts per renogram are of the order of 1,000 the optimum number of factors is two. A proposed method for estimating parameters of noisy curves by means of a factor table appears to be well suited for the generation of functional images of the kidneys, as it is accurate, efficient, and relatively insensitive to noise.

REFERENCES

1. H. H. Harman, Modern Factor Analysis, 3rd ed., The University of Chicago Press, Chicago, Ill., 1976.

2. R. Di Paola, C. Penel, J. P. Bazin and C. Berche, "Factor Analysis and Scintigraphy," p. 91 in Proceedings of the Fourth International Conference on Information Processing in Scintigraphy (Orsay, France, July 15-19, 1975), C. Raynaud and A. Todd-Pokropek, Editors.

3. P. Schmidlin and F. Rösel, "Application of Factor Analysis for Scintigraphic Picture Processing," p. 80 in Proceedings of the Fourth International Conference on Information Processing in Scintigraphy (Orsay, France, July 15-19, 1975), C. Raynaud and A. Todd-Pokropek, Editors.

4. P. Schmidlin, J. Clorius, R. Kubesch and K. Dreikorn, "Evaluation of Dynamic Studies By Means of Factor Analysis," in Medical Radionuclide Imaging, 1976 (Proceedings of a Symposium held in Los Angeles, USA, Oct. 25-29, 1976), IAEA, Vienna, in print.

5. M. J. Gelfand, F. H. Allen, M. V. Green, G. S. Johnston and J. J. Bailey, "Functional Mapping of Renal Transit of ^{131}I-iodohippuran," p. 100 in Proceedings of Third Symposium on Sharing of Computer Programs and Technology in Nuclear Medicine, USAEC CONF-730627, 1973.

6. J. A. DeGrazia, P. O. Scheibe, P. E. Jackson, Z. J. Lucas, W. R. Fair, J. M. Vogel and L. J. Blumin, "Clinical Applications of a Kinetic Model of Hippurate Distribution and Renal Clearance," J. Nucl. Med. 15, 102 (1974).

7. E. Knudsen and K. S. Hojberg, "Simulation of the Radiohippuran Renogram Test on an Analogue Computer," Int. J. Appl. Radiat. Isot. 18, 639 (1967).

8. A. Hirakawa, M. Kuwahara and H. Ueyama, "Analog Computer-Aided RI Renogram Diagnosis," p. 303 in Radionuclides in Nephrology, M. D. Blaufox and J. Funck-Brentano, Editors, Grune and Stratton, New York, 1972.

9. D. M. Shames, R. E. L. Farmer, R. R. Wright, F. D. Rollo, S. J. Swann and O. Salvatierra, "A Computerized Technique for Evaluating Renal Transplant Function," p. 209 in Proceedings of Fifth Symposium on Sharing of Computer Programs and Technology in Nuclear Medicine, USERDA CONF-750124, 1975.

10. J. D. Horgan, R. C. Meade, J. A. Madden and T. A. Torzala, "Digital Computer Simulation Study of the Radio-Hippuran Renogram," Int. J. Appl. Radiat. Isot. 18, 797 (1967).

11. T. Heiskanen, T. Weber and R. Grösbeck, "A Method for the Quantitative Evaluation of Radioisotope Renograms," Int. J. Appl. Radiat. Isot. 19, 827 (1968).

12. B. E. Oppenheim, "Three-Dimensional Reconstruction from Incomplete Projections," p. 288 in Proceedings of the Fourth International Conference on Information Processing in Scintigraphy (Orsay, France, July 15-19, 1975), C. Raynaud and A. Todd-Pokropek, Editors.

13. C. E. Metz, personal communication.

14. W. E. Adam, R. Kadatz, F. Bitter, E. Sigmund and H. O. Wack, "Investigations on Kidney Perfusion Tests with Radioactive Substances," p. 341 in Radionuclides in Nephrology, M. D. Blaufox and J. Funck-Brentano, Editors, Grune and Stratton, New York, 1972.

15. H. Agress, S. M. Levenson, M. J. Gelfand, M. V. Green, J. J. Bailey and G. S. Johnston, "Application of Computer-Generated Functional (Parametric) Maps in Radionuclide Renography," p. 180 in Proceedings of Fifth Symposium on Sharing of Computer Programs and Technology in Nuclear Medicine, USERDA CONF-750124, 1975.

16. R. E. Linnemann, M. K. Loken and C. Markland, "Computerized Compartmental Renograms to Study Kidney Function," J. Urol. 103, 533 (1970).

17. Y. Wang, "Regional (Compartmental) Renogram for Hypertension Evaluation," Radiology 118, 842 (1973).

18. K. zum Winkel, H. Jost, F. Motzkus and G. Golde, "Renal Function Studies with Radioisotopes," p. 229 in Dynamic Studies with Radioisotopes in Medicine, IAEA, Vienna, 1971.

19. L. Pöyhönen, H. Holli, "Etude, à l'Aide d'Examens Gammagraphiques et de Nephrographies Partielle à l'ADTP-^{169}Yb, des Modifications de la Circulation Arterielle des Reins Provoquant l'Ischemie," p. 275 in Medical Radioisotope Scintigraphy 1972, IAEA, Vienna, 1973.

20. A. Halko, G. Burke, A. Sorkin and J. Enenstein, "Computer-Aided Statistical Analysis of the Scintillation Camera ^{131}I-Hippuran Renogram," J. Nucl. Med. 14, 253 (1973).

DISCUSSION:

Pizer asked Oppenheim if he used 16 element vectors to describe his renogram curves rather than 64 or 128 elements and asked further if the size of the renogram vector affected the analysis.

Oppenheim stated that he felt that most of the information in a Hippuran renogram is present in the low frequency data components. Compression to 16 elements effectively reduces the contribution of the high frequency information. The high frequency component, however, is probably obscured by noise, anyhow, so he did not feel that much of the usable information is lost. Dr. Schmidlin has shown that the values of the main two factors stay relatively stable when the number of elements in the renogram is increased from 16 to 64 or more.

Britton stated that the most important use of this type of classification technique is in determining which regions of an image are background and which are not. When this technique is used to describe the function of the kidney on a pixel by pixel basis, however, he questioned the physical significance of the results. In an organ which is composed of tubes running in all directions and in which vessels, nephrons, and collecting tubules may all be included in a single pixel, it is clear that the information in any small region is very difficult to interpret.

Oppenheim responded by saying that he was led to this area because the generation of regions of interest over the kidney and background is a very tedious process, and he was hopeful that factor analysis would provide a means to automate region of interest selection. He stated that he assumed that these parameters had no physical meaning but thought they should allow one to separate background from kidney and perhaps renal pelvis from cortical regions. This technique should also be valuable for producing functional images from tables of factor versus functional parameter relationships.

CAROTID AND CEREBRAL BLOOD FLOW

K.E. Britton, C.C. Nimmon, T.Y. Lee, P.H. Jarritt
M. Granowska, A. Greening and Joan M. McAlister
Department of Nuclear Medicine and Radioisotopes
St. Bartholomew's Hospital
West Smithfield, London EC1, U.K.

ABSTRACT

With the aim of developing quantitative methods to aid the assessment and prevention of stroke, a new non-invasive technique using ^{99m}Tc albumin for the simultaneous assessment of the character of carotid artery flow and the measurement of regional cerebral flow is presented. The practical, theoretical and experimental basis for distinguishing between non-turbulent and turbulent flow in carotid vessels is described. The method depends on deconvolution analysis and estimates of errors are given. Regional cerebral flow in millilitres per minute is measured by dividing regional cerebral blood volume by regional mean transit time, both being assessed independently. The advantage of this technique over the ^{133}Xenon methods are described and useful clinical results have been obtained.

INTRODUCTION

The incapacitating effects of a stroke on a patient and his immediate associates makes the search for any aid to the prevention of strokes, an essential community and medical priority. Progress in the understanding and management of transient ischaemic attacks, a forerunner of many strokes has been hampered by the lack of reliable quantitative non-invasive methods for screening patients with suspected cerebrovascular disease. The main aetiology of many transient ischaemic attacks is embolism of fragments from atheromatously narrowed carotid and vertebral arteries. Such narrowings of these vessels are due to atheroma and evidence favours the concept that turbulence of flow in such vessels may be both atherogenic and embologenic. For example, the bifurcations of vessels are sites of turbulence and the usual site for atheroma. Turbulent flow is associated with eddy formation which helps to deposit material on or allow diffusion into a vessel wall. Turbulent flow is also associated with vortices which can contribute to the excoriation of deposited material from a vessel wall. During twisting movements

of the head and neck it is likely that the relative sites of eddies and vortices will alter.

Brice et al.[1] demonstrated in man that a significant pressure drop in a carotid artery only occurs with a 'critical' stenosis, that is one with an 80% or greater degree of stenosis. Carotid stenoses, having less than this haemodynamically critical degree of narrowing, are common, as are transient ischaemic attacks. It is postulated therefore that sub critical stenosis may induce turbulence and that this is the critical change that leads to the transient ischaemic attack. It is further postulated that detection of changes in the character of flow in major vessels such as those due to turbulence would be useful in following the natural history particularly in high risk groups e.g. young hypertensives and the effectiveness or otherwise of therapy. An attempt has been made to develop a non-invasive technique called carotigraphy[2] for the objective assessment of the character of carotid flow.

It is evident that whatever the changes in carotid and vertebral flow that are demonstrated, their effects are somewhat reduced by the presence of a functioning circle of Willis. Since this gives rise to three major arteries to each hemisphere, it would appear to be appropriate to combine with carotigraphy, a non invasive measurement of cerebral blood flow to these six regions. This paper reports the first quantitative combined measurements of carotid and cerebral flow. The paper outlines the progress so far, the technical pitfalls and the problems of interpreting the results in a clinical context.

Whereas cerebral blood flow - in fact cerebral perfusion in millilitres per minute per 100 grams of tissue - has mainly been measured using inert radioactive gas techniques with either their intraarterial injection[3], inhalation[4] or less reliably intravenous injection[5], our technique is designed for the use of a non diffusible indicator to give measurements of actual regional cerebral flow in millilitres per minute. Apart from the practical problems of the ^{133}Xenon washout technique, there are certain important theoretical objections[6]. The principles of Xenon^{133} washout are outlined below :

Cerebral blood flow = Volume . Rate of washout, λ

$$CBF = V \cdot \lambda$$

V = Volume of distribution.

Since V depends on the partition of Xenon between tissue and blood, then $V_{Xenon} = V \cdot P$

P = Partition coefficient for Xenon

Using Xenon

$$CBF/\ ml/min = V \cdot P \cdot \lambda\ ml/min$$

Cerebral Perfusion, CP = CBF/unit weight of tissue, w

$$CP = \frac{CBF}{w} = \frac{V \cdot P \cdot \lambda}{w}\ ml/min/gram$$

Now w/V = density ρ grams/ml

$$CP = \frac{P \cdot \lambda}{\rho} = P \cdot \lambda, \text{ for } \rho = 1$$

From this formulation it can be seen that firstly although the actual measurement is of λ whose units are Sec^{-1}, by the introduction of density and setting its value to unity the extra dimensions of ml. $gram^{-1}$ are introduced. Secondly since $CP = P \times \lambda$, an assumption that P is constant has to be made for CP to be directly proportional to λ. In health, this may be true for white and grey matter but there is no evidence to show that it holds true in disease. Thirdly it is assumed that the volume of distribution of Xenon is not only constant but occurs virtually instantaneously through the whole brain, yet it is not at all clear how much Xenon actually enters grey and white matter and whether the amount in neuroglia differs or not from that in nerve tissue. The assumption of free diffusibility of Xenon into brain appears to be made on the basis of the observed free diffusibility of Xenon from skin and muscle. As Olendorf[7] has re-emphasised[7], the structure of the endothelial lining of cerebral capillaries is quite different from that of other capillaries. The blood brain barrier is both a physical entity, in that cerebral capillaries do not have fenestrations or pores; and physiochemical entity with differing modes of transport for polar and non polar substances. These fundamental problems overshadow the numerous sources of systematic error of the $Xenon^{133}$ washout technique.

An alternative method for cerebral blood flow measurement has been implemented. It is based on the measurement of fractional cardiac output to the cerebral region of interest and this technique has essentially the assumptions inherent in determining cardiac output using an external detector system. These latter are minimised by maintaining the same geometrical detector-patient configuration during the first pass and the equilibrium phases of the measurement as with the measurement of cardiac output.

MATERIALS AND METHODS

Carotigraphy, Apparatus and Technique

In order to monitor the flow characteristics of the carotid arteries from the aortic root to the internal carotid in the cavernous sinus, a pair of externally placed scintillation detectors are positioned over the orbits collimated so as to focus on each internal carotid artery in the cavernous sinus[2,8]. These detectors were designed by M.J. Myers. Each has a 7.5 x 2.5 cm crystal and a coverging collimator, 7.6 cm diameter at the crystal face and 5 cm diameter at the front face. Each contains 37 tapered holes with 1 mm lead septa and has a focal point of 9 cm (Full width half maximum, FWHM, 14 mm). The crystal is shielded with 3 cm thick lead and the photomultiplier tube with 1 cm thick lead to exclude extraneous radiation when the injection is brought to the patient's arm. The detector assembly is mounted on one arm extending from a stand, and capable of vertical, horizontal and lateral adjustments as well as rotation along the horizontal and vertical axes. This allows the detect-

ors to be positioned over the patient's eyes with the focal points on each cavernous sinus. A third detector is mounted on a mobile stand and is positioned over the aortic root at the level of the second left intercostal space. It has a 5 x 2.5 cm crystal with a 4.5 cm long conical converging collimator, diameter 4.5 cm at the crystal face and 2.4 cm externally, having a central lead plug and 8 radiating vanes. Its focal point is 7 cm (FWHM 10 mm). Each detector feeds signals into standard international series electronic equipment feeding a chart recorder and a paper tape punch. Data is accumlated at 0.3 second intervals.

About 30 minutes after 400 mg potassium perchlorate orally, the patient reclines on a couch with its back rest elevated to about 45°. With the patient's arm abducted, 12-15 mCi technetium 99m labelled albumin is injected intravenously in a volume of less than 0.4 ml through a guage 16 cannula. The system is arranged so that the bolus is introduced and followed by 20 ml 0.9% saline in one swift movement. The carotigraphy traces on the chart recorder are analysed for the degree of asymmetry[2]. The paper tape is fed into a Varian 620/L100 computer and deconvolution analysis is performed.

Cerebral Volume, Mean Transit Times and Flow

During brain scanning with Technetium-99m pertechnetate given intravenously, many centres have used a gamma camera to follow the progress of the bolus of activity through the brain[9] as a prelude to static studies. Moses et al[10,11] favoured the vertex view in order to separate the anterior, middle and posterior cerebral artery distributions and obtained useful qualitative results. Quantitation of vertex view regional cerebral blood flow has been described by several authors[12,13,14,15]. None of these authors have made allowance for the effect of variations in carotid artery flow and variations in the shape of the input activity/time curves entering the brain. Goddard et al[16] partially corrected for this effect by monitoring the bolus of activity as it passed through the lung. By combining carotigraphy with vertex view regional cerebral flow measurements, we have been able to obtain by deconvolution analysis a rather rigorous solution to this problem. The gamma camera is set up over the vertex of the patient reclining at 45° and the special lead shielding is fitted snugly to the neck and over the shoulders.

The measurement of cardiac output using radionuclides is well accepted. A first pass cardiac activity/time curve is recorded by an externally placed detector and the curve is calibrated by counting a blood sample at equilibrium and relating its activity to the height of the activity/time curve at equilibrium. Flow in ml/min is then calculated as Volume/mean transit time. Using the same approach a computer-linked gamma camera is placed over the vertex of the patient's head and shielded from the upper trunk. First pass activity/time curves are recorded simultaneously from usually 6 selected regions on interest and from the whole vertex. An equilibrium image is recorded at 2 minutes for one minute. At 3 minutes a blood sample is obtained and then counted in a phantom under the camera.

A blood volume V_R, corresponding to each ROI (R) is calculated from the formula:-

$$V_R = \frac{N_R}{N_T} \cdot \frac{(E_q - E(t))}{S}$$

Where,

N_R is the total counts within the region in a summed count image formed by summation of the raw data over the period of the first pass of activity.

N_T is the total first pass activity.

E_q is the total counts in an equilibrium image recorded between 2 - 3 minutes post injection.

S is the value of the counts/ml/min corrected for geometry obtained from a plasma sample taken at 3 minutes post injection.

E (t) is the contribution to the equilibrium image from diffusion and is not truly zero for the ^{99m}Tc human serum albumin used as the tracer in this study. ^{99m}Tc human red cells will soon be used as a less diffusible and more appropriate alternative. E(t) may be estimated by the times series study or by a double radionuclide technique. E(t) will increase with blood brain barrier disruption and may be used as a measure of that disruption. Comparison of a metabolically active tracer and a non-diffusible tracer will lead to an estimate of metabolic brain activity separate from blood.

The true mean transit time (T_R) of each of six regions of interest (ROI's) defined on a vertex view is calculated as follows. Deconvolutions are carried out using the activity/time curve from the aortic probe as the input function and each ROI activity/time curve to give an impulse retention function. These functions are corrected for the transit times from the aortic probe to each carotid probe. The functions are then differentiated to obtain the transit time spectrum from which the mean transit time T_R may be calculated.

Before applying the deconvolution method, it is necessary to determine the appearance times of the input/output curves. The curve obtained from the aortic probe contains a contribution due to cross talk from activity in the superior vena cava and the lungs. The leading edge of this input curve is linearly extrapolated to remove this contribution[17]. A similar linear extrapolation is carried out to obtain the appearance time for the curves from the carotid probes and the vertex ROI's. A value for the appearance time on the vertex view is subtracted from the mean transit times obtained by deconvolution for each cerebral region to allow for transit between the aortic root and the base of the brain. An alternative method for this

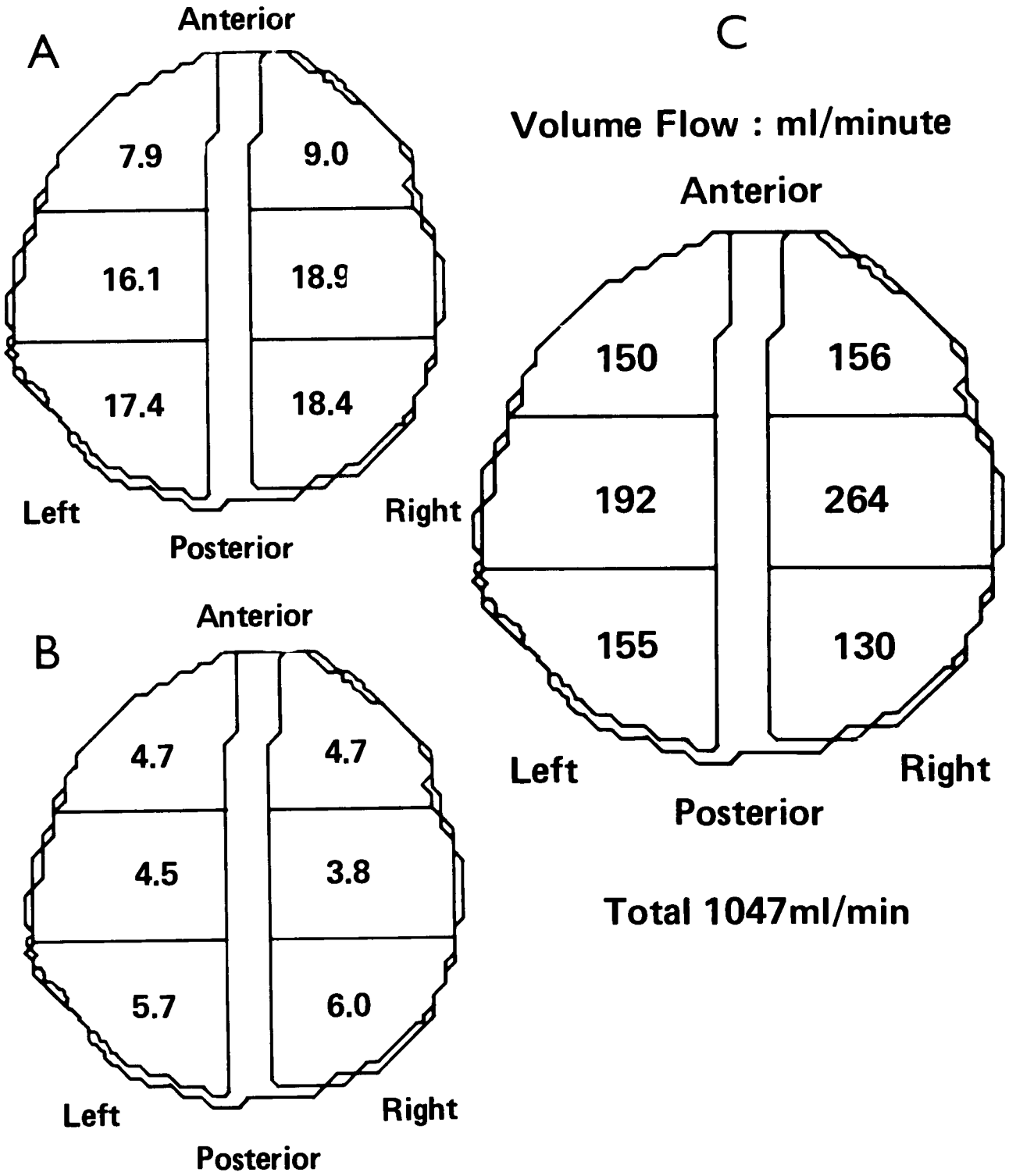


Fig. 1.
Diagrammatic representations of regional cerebral blood flow in a man with bilateral carotid stenosis left greater than right.
(a) Volume in ml
(b) Mean transit time in seconds
(c) Flow in ml/minute

correction would be to use the convolution of the first peak in the impulse response functions obtained for the carotid flow, and the aortic curve as the input function for the cerebral ROI's. However this would probably overestimate the width of the true function because of the theoretical effects discussed subsequently. The vertebral circulation constitutes a second input to the cerebral system. This may be expected to show as a second rise in the retention function. Any re-entry of tracer to a ROI will have the same effect. In calculating the mean transit time from the retention curves using equation (3), 'negative' transit times have therefore been ignored. Flow in ml/min through each region of interest (F_R) is then given by $F_R = V_R/T_R$. These flows are then corrected for the number of elements in each pair of regions, front, middle and posterior. Regional cerebral blood flow is obtained in millilitres per minute and the data is displayed in diagrammatic form (Figure 1).

A limitation of the quantitative techniques at present is imposed by the computer time required to carry out the deconvolutions on a Varian 620/L100 minicomputer. Approximately three hours per patient are needed and the calculations are carried out on an overnight batch run.

For an immediate qualitative assessment of the study, use is made of both a 'mean time' and an 'appearance time' functional image. (a) Mean time image. The mean time, of the vertex activity is calculated at each of the 64x64 pixels in the vertex view and displayed as the intensity variable.

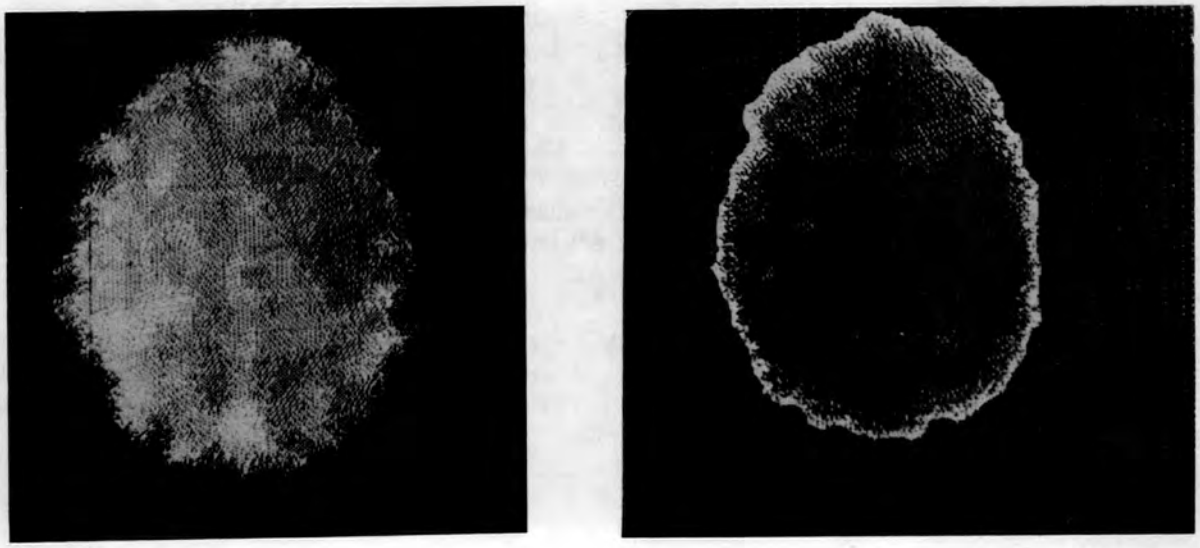
$$t_{mean} = \frac{\sum_i t_i \cdot N_i}{\sum_i N_i}$$

Where N_i is the number of events measured between the time t_i and t_{i-1}.

The percentage error on the mean time arising from statistical noise is also calculated at each pixel using standard error formula and displayed as the intensity variable in an 'error image' with highest intensity corresponding to an error of 10%.

$$\% \text{ error} = \left[\frac{\sum_i t_i^2 \cdot N_i}{\left(\sum_i t_i \cdot N_i\right)^2} + \frac{1}{\sum_i N_i} \right]^{\frac{1}{2}} \cdot 100$$

Within each ROI the statistical error on the mean time is found to be less than 7% except for the peripheral boundary regions where the error approaches 10%. Examples of a mean time and error image are shown in Fig. 2. For a given input, differences in mean time reflect differences in mean transit time and the example in Fig. 2 shows definite evidence of asymmetry between the transit times in the two hemispheres.
(b) Appearance time image.
For each pixel, an appearance time is calculated using the



A B

Fig. 2.
Vertex view results in a man with left carotid stenosis.
(a) Mean time image
(b) Error image

alogrithm:-

appearance time $= 2.t_j - t_k$

where t_j, t_k are the times required for the counts observed to exceed $\frac{1}{4}$, $\frac{1}{2}$ of the maximum counts respectively.

An example of an appearance time image is shown in Fig. 3 (d) and shows arrival of activity in the posterior before appearance in the anterior regions.

Application of Deconvolution Techniques to Blood Flow

The impulse response, g(t), for a system may be defined as the observed output from the system arising from a unit impulse input. The output from the system, O(t), arising from a non-spike input, I(t), is then given by the convolution equation:-

$$O(t) = \int_{c}^{t} I(\tau) \cdot g(t - \tau)\, d\tau \qquad (1)$$

Application of this equation involves three main basic assumptions:-

(a) Linearity of the system. (b) Stationarity. The impulse response must not vary during the period of the study. (c) The tracer must form a homogeneous mixture with the particles being traced at the point of observation of the input.

The amount of tracer contained within the system, R(t), is related to the input by a similar equation:-

$$R(t) = \int_{0}^{\tau} I(t) \cdot G(t - \tau)\, d\tau \qquad (2)$$

where G(t) is defined as the impulse retention function of the system. This retention function is related to the impulse response function by the equation:-

$$g(t) = - \frac{dG(t)}{dt} \qquad (3)$$

If the input along with either the output or the system content are known, then deconvolution techniques can be applied to determine the impulse response from equations (1), (2), (3). The impulse response is a characteristic of the system under study. In blood flow studies the impulse response may further be related to basic parameters of flow such as the mean transit time and mean flow rate. The relationship between these quantities will depend on the type of flow predominating.

For the situation in which the tracer remains adequately mixed with the fluid between the input and output regions,

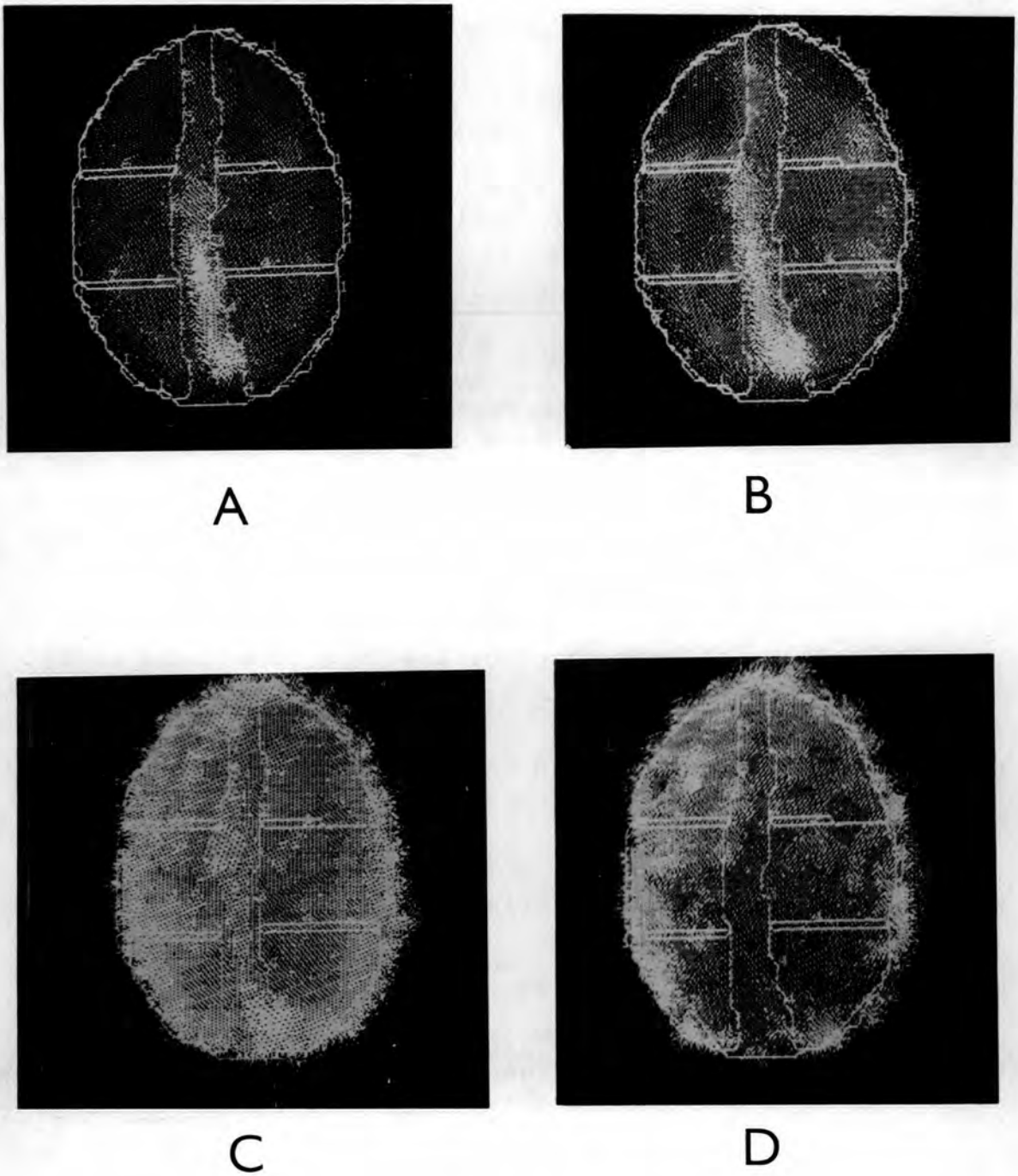


Fig. 3.
Vertex view result in a man with right hemiparesis see also Table VI.
(a) Summed count image
(b) Equilibrium image
(c) Mean time image
(d) Appearance time image

the impulse response is numerically equal to the frequency distribution of transit times, h(t). The mean transit time, $\bar{t}$, of the system is given by:-

$$\bar{t} = \int_0^\infty h(t) \,.\, t\, dt = \int_0^\infty g(t) \,.\, t\, dt \qquad (4)$$

and is related to the mean flow rate, F, and the volume of distribution of the tracer, V, by the basic equation of indicator dilution theory[18].

$$F = V / \bar{t} \qquad (5)$$

This equation applies both to flow in a single vessel and in a vascular bed and will also cope with recirculation within the system. The application of deconvolution techniques to cerebral blood flow in this study and the calculation of mean transit times and flow rates depends on the validity of the 'mixing' assumption. Effective mixing is considered likely to occur on account of the proliferative degree of branching and the complex geometry of the vessels encountered in the cerebral circulation.

However, in the major arteries between the aortic root and the base of the brain, blood flow is thought to be predominantly of the laminar flow type. Deviations from a classic Poiseuille form will of course undoubtedly occur, for example a blunting of the velocity profile due to the pulsatile nature of the flow and a localised degree of mixing at the branching of the carotid vessels.

For mean flow rates of the order of 20 cm/sec in a straight tube, radial dispersion of indicator by diffusion[19] can be neglected[19]. Thus for the carotid flow studies it is therefore necessary to consider further the consequences of incomplete mixing of indicator across the vessel and in particular to determine the relationship between the impulse response and the transit time spectrum.

The frequency distribution of tracer transit times, h(t), is defined as the flux of tracer particles crossing a plane at the output region following a homogeneous unit impulse input flux. For laminar flow the velocity at a radial distance r is given by the equation:-

$$v(r) = u. \left[1 - \left(\frac{r}{a}\right)^n \right] \qquad (6)$$

where u is the maximum axial velocity,
a is the radius of the vessel,
and n is the effective velocity profile number and is equal to 2 for a parabolic velocity profile.

Following a unit spike input, the indicator observed in the output region will be contained within the range $r_1(t) \leqslant r \leqslant r_2(t)$.
The flux is given by:-

$$h(t) = \int_{r_1}^{r_2} 2\pi r \cdot C(r) \cdot v(r) \, dr \qquad (7)$$

where C(r) is the radial concentration of indicator in the output region. For laminar flow, the concentration within the leading and trailing edges of the indicator profile remains constant, i.e.

$$C(r) = C \quad \text{for} \quad r_1(t) \leqslant r \leqslant r_2(t)$$

$$= 0 \quad \text{for} \quad r < r_1(t) \ , \ r > r_2(t)$$

A mean velocity of the indicator particles observed within the output region following a unit spike input is defined by the euation:-

$$\bar{v}(t) = \frac{\int_{r_1}^{r_2} r \cdot v(r) \, dr}{\int_{r_1}^{r_2} r \, dr}$$

The amount of indicator within the output region, A(t), is given by:-

$$A(t) = \int_{r_1}^{r_2} 2\pi r \cdot C \cdot \Delta w \, dr$$

where Δw is the width of the observation region.

Hence, from equation (7),

$$h(t) = \frac{A(t) \cdot \bar{v}(t)}{\Delta w} \qquad (8)$$

where A(t) is the observed amount of indicator in the output region following a unit input flux. If complete mixing of indicator across the vessel cross section is assumed at the input region, the mean velocity of the particles observed there is equal to the mean flow velocity, F. For a unit amount of indicator at the input the observed amount at the output, the impulse response, is then given by:-

$$g(t) = \frac{F \cdot h(t)}{\bar{v}(t)} \qquad (9)$$

where the width of the observation region at the input has been taken equal to that at the output. (Note that for complete mixing between the input and output regions $\bar{v}(t)=F$ and $g(t)=h(t)$). Hence for laminar flow, deconvolution of the activity/time (A/T) curves does not yield the transit time spectrum h(t)

directly. However for a straight vessel of length l, $\bar{v} = l/t$, and the time dependence of h(t) may easily be determined. An example of the functions g,h calculated for a parabolic velocity profile, a mean flow velocity of 20 cm/sec and a vessel length of 40 cm, is shown in Fig. 4a.

For flow in a straight vessel, the expression for the mean transit time becomes -

$$\bar{t} = \frac{\int_0^\infty g(t) \quad dt}{\int_0^\infty g(t) \cdot t^{-1} \, dt} \qquad (10)$$

This formula has been suggested by Lane and Sirs[20] and is applicable for a general form of the velocity profile. Equation (5) is no longer valid due to imcomplete mixing of the tracer between the input and output regions. The expression for the mean flow velocity F may be derived from equation (6) and is given by:

$$F = \frac{n}{n+2} \cdot u = \frac{n}{n+2} \cdot \frac{l}{t_a} \qquad (11)$$

where t_a is the appearance time of the tracer at the output.

A further difficulty in interpreting the impulse response obtained by the deconvolution technique arises if the tracer is not homogeneously mixed across the vessel cross section at the input observation site, for example in the case where laminar flow has already been established. The convolution equation in this situation may be written as:-

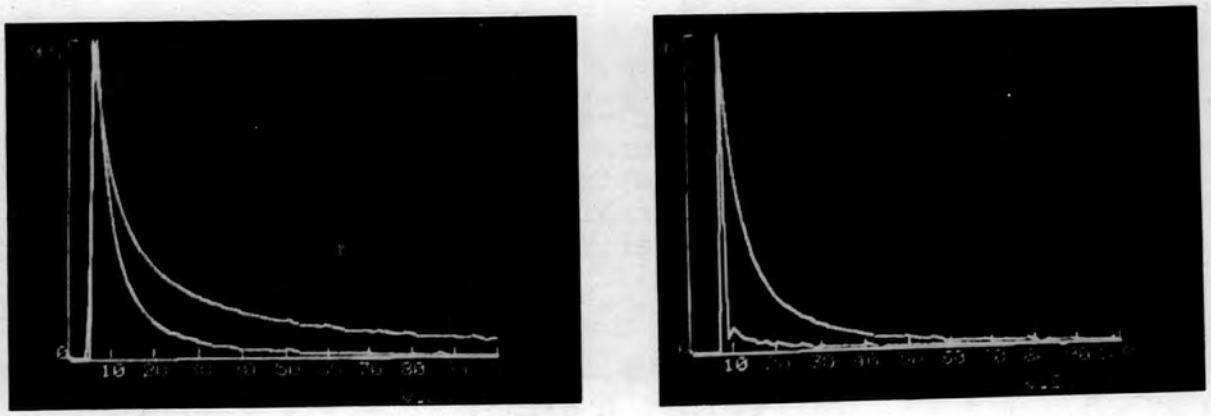
$$O(t) = \int_{\tau=0}^{t} \int_j I_j(\tau) \cdot g_j(t-\tau) \, dj \, d\tau \qquad (12)$$

$$\text{and } g(t) = \int_j g_j(t) \, dj \qquad (13)$$

where $I_j(t)$ and $g_j(t)$ are the input and impulse response respectively for a laminar flow layer j. The mean velocity of particles in each layer j is the same at the input and output and the velocity ratio is unity, i.e. $h_j(t) = g_j(t)$.

In practice, using radionuclide techniques, it is not possible to determine the variation of the input with radial position as only the mean concentration is recorded. Application of equation (1) will lead to a pseudo-transit time spectrum. An example of such a spectrum has been calculated theoretically and is shown together with the true transit time spectrum in Fig. 4b. The main characteristic of the pseudo-spectrum is that it consists almost entirely of a single very sharp peak.

Results of deconvolution of the A/T curves obtained in vivo



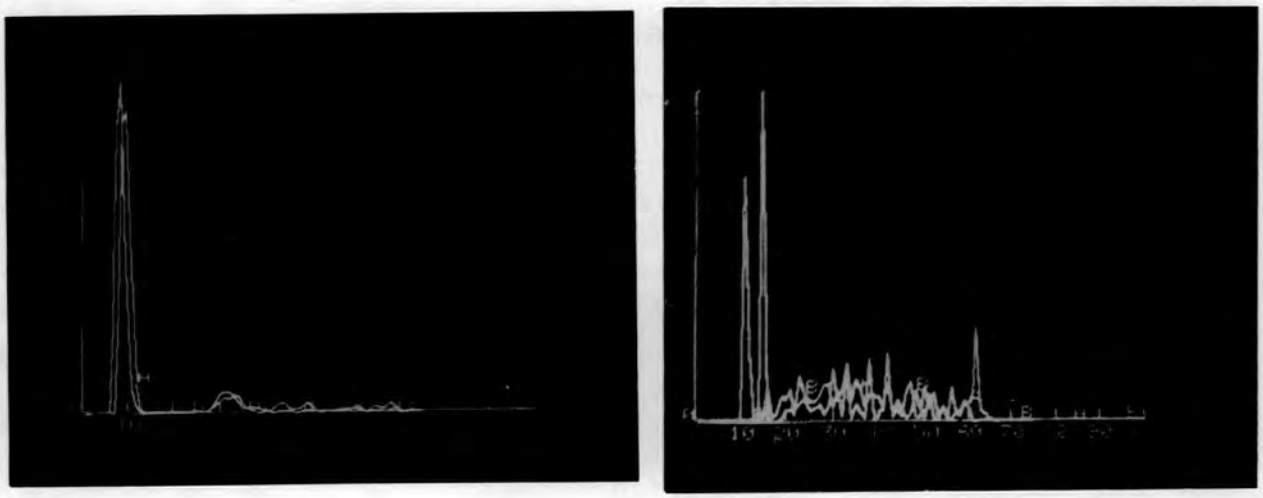
A B

Fig. 4.
Theoretical curves for laminar flow with a parabolic velocity profile in a straight vessel of length 40 cm and a mean velocity of 20 cm/sec.
(a) Functions g(t) (Upper curve) and h(t)
(b) Functions h(t) (Upper curve) and the pseudo-transit time spectrum

from the aortic and carotid probes closely resemble this narrow shape. Examples are shown in Fig. 5. Only the first peak of the spectrum can be clearly identified with carotid flow as later contributions may be due to circulation of the tracer in the cerebrum in the field of view of the detectors. Preliminary calculations using a comparison of the width at half maximum suggest that these curves may be simulated by choosing a value of n, the effective velocity profile number, in the range 3 - 4. An explanation of the narrowness of these curves in terms of a broad velocity profile is thought to be more probable than an alternative conjecture of a deviation from homogeneous mixing in the region of the aortic arch. The determination of an effective velocity profile number may in future provide a convenient method of classifying these curves.

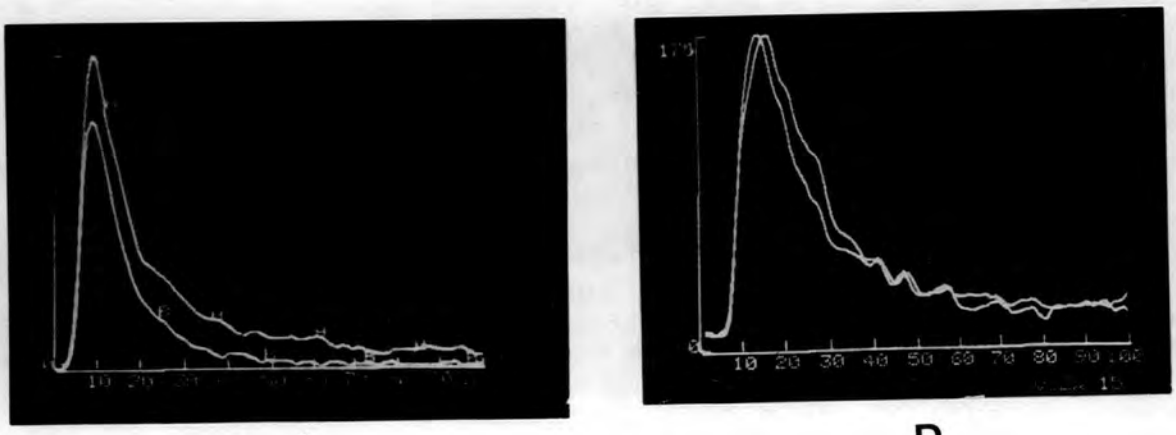
The basic aim of the carotid flow studies is to detect alteration to normal flow patterns and in particular, the occurrence of turbulence. The deconvolution technique in removing the dependence of the output curve on the input and in separating the cross talk from activity in the cerebrum, yields the impulse response which should be more sensitive in reflecting this effect than the original activity/time curves. Turbulence in laminar flow can be expected to have two main effects. Firstly, indicator near the walls of the vessel having a low velocity may be transported to the central high velocity region. Secondly, the converse effect causing particles to move from the central high velocity region towards the peripheral low velocity area. In either case, the non-uniformity of flow velocity would be reduced. The detailed shape of the impulse response obtained distal to the site of the localised turbulence may be expected to depend on which of these mechanisms predominates. These effects have been demonstrated experimentally with a blood flow rate of 20 cm/sec in a straight plastic-walled tube of diameter 0.7 cm. Two types of constrictions producing turbulence were used:- (A) Cylindrical occluding 85% of the vessel, (B) A projection of 1 mm. A/T curves were obtained from sites on both sides of the constriction, and the results given by averaging the curves from five experiments carried out under identical flow conditions are shown in Fig. 6.

(A) The curve obtained distal to the constriction is consistent with streaming of the indicator through the remaining annulus, i.e. increased velocity components appear dominant. (B) The corresponding curve is consistent with some disturbance of the highest velocity components. The effect of increasing the uniformity of the flow velocity may be expected to reduce the variance of the observed curves. The observation of reduced variance and increased appearance times associated with mixing effects in predominantly laminar flow in 'in vitro' experiments has been reported by Caro[21]. This effect of decreased variance caused by turbulence in laminar flow conditions is opposite to the expected effect in predominantly mixed flow conditions. Here, turbulence acting as a randomising effect might theoretically, applying a random walk model, be expected to increase the variance. Thus the resulting effect is likely to depend on the degree and extent of the turbulence as well as the predominant flow mode. The degree of turbulence might in turn be expected to vary with the nature of the projection producing



A B

Fig. 5.
Examples of the impulse response functions g(t) obtained from 'in vivo' carotid flow studies (The scale of the time axis is in units of 0.3 sec)



A B

Fig. 6.
A/T curves obtained in 'in vitro' experiments following a bolus injection of indicator. Mean flow velocity = 20 cm/sec. Curves shown correspond to positions on either side of a constriction and were obtained by averaging the results of five experiments.

it. It could be postulated that a smooth projection, e.g. a pure atheromatous plaque would cause a narrowing of the impulse response, whereas a plaque which had ulcerated with a clot attached would lead to a greater degree of dispersion and be demonstrated as a widening of the impulse response.

Clearly in order to test these hypotheses, further in vitro experiments are needed to simulate these effects under controlled conditions of flow. The ultimate value of the carotid flow studies will only be capable of assessment after a long series of patient studies and correlation with evidence obtained at surgery.

Deconvolution Method

The method used in this study is of the Point-by-Point type in which the impulse response is not constrained to take any preconceived form apart from the exclusion of negative values. This type of method was preferred by Gamel et al[22] in a comparative study of deconvolution methods as applied to the determination of the impulse response from indicator-dilution curves.

The particular iterative technique used here is based on a Monte Carlo principle. The function g(t) is expressed as a set of estimated values Y(n) along with a set of estimated errors E(n), where

$$n = 1,2.....,m$$

$$= 1 + t - t_a$$

and t_a is the appearance time of the tracer at the observation point of the output curve. The method consists of a search for better estimates Y(n) consistent with a reduction of the Chisquared value formed from the reconvoluted output and the observed output. By definition, each individual value Y(n), must lie within the interval $0 \leqslant Y(n) \leqslant 1$. Each estimate, Y(n), is chosen as the most probable value of the random variable x in a theoretical sampling distribution $S_n(x)$, defined in the interval $0 \leqslant x \leqslant 1$. The normal distribution is chosen as the form of S (x) and the value Y(n) is then the mean (μ) of this distribution and E(n) is chosen as the standard deviation (σ).

A pseudo-random number generator is used to sample the distributions $S_n(x)$. The current sample values are then used in the calculation of the current chisquared value. A value lower than the previous value causes the sample value to replace Y(n) and the distribution $S_n(x)$ is moved so as to be centred about the new value. A higher chisquared value causes no change in the position of $S_n(x)$. Each individual distribution $S_n(x)$ continues to be sampled as long as new estimates of the particular Y(n) are being obtained. Five successive failures to obtain a better estimate causes the program to proceed to the next distribution ($n \rightarrow n+1$). At the end of each cycle, the current values of E(n) are reduced by a factor 0.5 within the range $0.8 \leqslant E(n) \leqslant 0.01$. Also after each cycle a simple (1,2,1) smoothing of Y(n) is carried out and the resulting values subjected to the above chisquared test. This is an empirical attempt to take into account the correlation between the Y(n) values.

The current value of chisquared is displayed on the visual display unit of the computer and the program can be terminated automatically after a fixed number of cycles or by operator intervention.

It has been pointed out[22] that in the presence of statistical noise, comparison of the reconvoluted output with the actual output is not a sufficient guarantee of the accuracy of the determined impulse response. It is necessary to carry out noise simulation studies to obtain a measure of the accuracy of any particular deconvolution method.

To carry out such checks, a theoretical input and impulse response curve have been chosen of a form resembling the experimentally encountered functions.

Since the aortic probe curves have high statistical accuracy compared with those obtained for each cerebral ROI, noise has been added only to the output curve. A pseudo-random number generator was used to simulate statistical noise using a gaussian distribution with width proportional to the square root of the value of the curve at each point. Varying noise levels were obtained by altering the total area, A, under the output curve. The output curve had 90 data points and the true impulse response had 8 points. Deconvolutions were performed for a specified maximum number of parameters m = 8,9 .....,20.

The accuracy of the resulting impulse response has been assessed in terms of a percentage error in each of the moments - mean, variance, skewness - and also by calculation of a chisquared value formed between the derived and theoretical impulse response. The results are summarised in Table 1.

Table 1. Results of Tests of the Monte Carlo Deconvolution Technique using Simulated Statistical Noise

Area A (counts)	Chisquared/ Deg. of Freedom		Mean		% Error on Variance		Skewness	
	m=9,	m=20,	m=9,	m=20,	m=9,	m=20,	m=9,	m=20
1.10^5	8.2	9.0	-0.6	-2.1	+18.7	+43.6	+188.2	+460.7
5.10^4	30.2	7.9	-1.5	-3.1	+15.6	+33.2	+244.7	+430.1
2.10^4	17.1	4.6	-3.5	-4.0	+11.4	+32.7	+194.7	+300.0
1.10^4	16.2	8.0	-5.2	-6.5	+ 9.2	+29.3	+ 70.8	+378.9
5.10^3	8.0	39.3	-7.5	-7.7	+20.1	+60.4	+277.1	+567.3
2.10^3	28.0	64.3	-21.4	-27.0	-27.3	-12.8	- 81.5	+ 34.87
1.10^3	39.5	82.9	-26.6	-28.9	-36.8	-29.5	- 68.5	- 16.60
5.10^2	104.8	165.0	-42.4	-43.5	-60.3	-45.8	- 90.1	- 34.2

The χ^2 test shows good agreement for $A \geqslant 10^4$ (maximum height of output curve = 273.) In the range $A \geqslant 10^4$, the error on$(\bar{t} - t_a)$ is less than 7%. The error on $\bar{t}$ is determined mainly by the error in calculating t_a. Study of the results for the variance, V, and skewness, S, shows a dependence on the number of 'redundant' parameters $m > 8$ in the spectrum. In practice for the carotid studies, it is probably possible to determine m visually to within a couple of points before calculating S,V. These results of the simulation studies enable errors on the derived impulse response to be estimated from a knowledge of the level of statistical noise present on the original data. In practice, all the A/T curves in the study have an area $\gg 10^4$.

CLINICAL RESULTS

57 patients with cerebrovascular disease were studied. In the first series of 27, only an index of the regional cerebral blood flow was estimated. The mean time of the first pass activity/time curve was used instead of the mean transit time of the region of interest; and an index of volume was obtained from the equilibrium image. This series, as well as giving semi-quantitive and qualitative results, helped to establish the feasibility of the technique, to improve the positioning of the patient relative to the detectors and to test changes made to the lead collar which shields the rest of the body from the gamma camera positioned over the vertex.

The results presented here concern 28 ambulatory patients with moderate cerebrovascular disease. In this second series the full technique was applied. Six of these patients were studied with 10-15 mCi ^{99m}Tc sodium pertechnetate, with one technically poor result. Results for volume and flow are given in Table II. The other twenty two patients were injected ^{99m}Tc human serum albumin as a less diffusible indicator. Some results were not obtained in 3 patients due to technical problems and this accounts for the variation in numbers of patients given in Tables II-V.

As examples of the clinical usage, two case histories are briefly described. A man of 56 with moderate hypertension under medical control had a left hemiparesis of sudden onset which recovered within 24 hours. His cerebral flows were measured two days later. He was well for nine weeks and then collapsed with a severe right hemiparesis with dysphasia. His cerebral flows were re-measured a few days later. These results are shown in Table VI. The changes in regional cerebral flows are appropriate to the clinical findings.

A man of 47 developed a subarachnoid haemorrhage without lateralising or focal neurological signs. Carotid angiography showed an anterior communicating aneurysm. After this he developed a severe right hemiparesis. Cerebral flow studies were undertaken two days later to see if his regional and total cerebral flows were reduced, Table VII. The reduction in left frontal and middle flow is seen in the presence of a normal total flow.

Table II. Cerebral flow: Comparison of Pertechnetate with Technetium Albumin

Type of Injection	No.	Volume ml Mean	Range	No.	Flow ml/min Mean	Range	Mean Transit Time sec
99m Tc TcO_4	6	257	155-408	4	3996	2956-4845	3.859
99m Tc albumin	22	141	99-256	19	2578	1209-4111	3.282

Table III. *Estimate of 'Normal' Values from Patients with Cerebrovascular Disease

	Side	No.	Mean	Range	Side	No.	Mean	Range	Total	Mean
Frontal	Right	14	313	184-439	Left	14	318	157-664	R+L	631
Middle	Right	7	749	366-1044	Left	7	725	349-887	R+L	1474
Posterior	Right	12	418	210-646	Left	12	433	251-710	R+L	851
Total	Right	33	1480 ml/min		Left	33	1476 ml/min		R+L	2956

* Patients with total cerebral flow greater than 1500 ml/min. Left/(Left + Right) value is within 40-60% for each region, 99m Tc albumin only.

Table IV. Asymmetry of Regional Cerebral Flow in 19 Patients with Cerebrovascular Disease

	Number of patients	Range of variation Flow ml/min	% asymmetry*
Frontal	2/19	69- 861	± 11-19%
Middle	10/19	135-1044	± 11-27%
Posterior	7/19	61- 651	± 12-21%

* The range of regional variation is expressed as (flow on one side as a percentage of total) which is outside the normal range 50% ± 10%, taking ^{99m}Tc albumin studies only.

Table V. Carotid Flow Characteristics in Patients with Cerebrovascular Disease

Type of injection	No of patients	Range of values 0.3 sec recording time variance	skewness	kurtosis	Mean transit time (secs)
$^{99m}Tc\ TcO_4$	5	L 0.486-4.381	0.056-2.385	0.486-20.33	6.44
		R 0.499-1.018	0.015-0.944	0.100-3.086	5.63
^{99m}Tc Albumin	21	L 0.253-4.702	0.005-5.863	0.075-52.86	4.85
		R 0.423-4.506	0.005-8.260	0.133-54.53	4.66
Order of variation		2 - 20 times	6 - 1000 times	30 - 700 times	

Table VI. Change in Flow with two Cerebrovascular Episodes in the same Patient

			Left transient hemiparesis 8.3.77	Right hemiparesis 17.5.77
Mean transit time ROI Secs	Frontal	L	1.4	5.2
		R	1.6	4.3
	Middle	L	1.1	4.3
		R	2.1	4.6
	Posterior	L	1.5	4.1
		R	2.7	2.7
Volume of ROI ml	Frontal	L	6	15
		R	10	20
	Middle	L	14	23
		R	18	30
	Posterior	L	14	26
		R	15	19
Regional flow ml/min	Frontal	L	297	185
		R	304	262
	Middle	L	900	349
		R	455	366
	Posterior	L	587	326
		R	299	420

Table VII. Regional Cerebral Flow after Subarachnoid Haemorrhage with Right Hemiparesis

		Mean transit time sec	Volume ml	Flow ml/min	Flow % Total
Frontal	L	3.25	15.2	280	35
	R	2.44	15.1	372	
Middle	L	3.18	30.7	580	33
	R	1.50	29.8	1190	
Posterior	L	4.73	26.0	330	47
	R	4.09	24.9	365	
Total	L + R		141	3087	

DISCUSSION

From the multitude of studies with inert gases of total cerebral perfusion a value of 55 ml/min/100 grams of tissue has been given as normal[24]. For 1500 gram brain, this gives the rather low value of 825ml/min whereas cerebral blood flow is usually thought of as being about 20% of a cardiac output e.g. 1200 ml/min. Our results with Tc albumin of 2956 ml/min flow, Table III, and with Sodium pertechnetate of 3996 ml/min, Table II are evidently high. However they cannot be said to be representative of the healthy middle aged population since they are taken from ambulant patients with known cerebrovascular disease. In such patients the blood brain barrier may well not be intact even for Tc albumin. A mean volume of distribution for sodium pertechnetate of 257 ml and a mean cerebral blood volume of 141 ml were obtained, Table II. These values include the sagittal sinus. These may be compared with the value of 5% of brain weight given by Zilkha[25] evaluated using an xray CAT scanner technique. This value is equivalent to 75 ml for a 1500 gram brain. Work is in progress using red cells firmly labelled with ^{99m}Tc and we expect its mean transit time and the measured volume of distribution to be less than with Tc albumin but not as low as 75 ml.

A number of variations are not yet accounted for in our technique. These include corrections for haemotocrit, carbon dioxide tension, and certain differences in geometry between the first pass and equilibrium states, as found in the measurement of cardiac output. By the use of a head band, some reduction in extra cranial activity may be obtained;[23] but the effect of this activity is partly cancelled out by the use of the same geometrical configuration for the first pass and equilibrium measurements. These factors will be evaluated in a later series of patients. The use of deconvolution analysis and the aortic and internal carotid probes does correct for the nature of

the input to the brain and for recirculation. The separate measurement of volume of distribution and mean transit time gives a more rigorous approach to regional and total cerebral blood flow then that attainable by the inert gas washout techniques.

The degree of asymmetry in the presence of local cerebrovascular disease is summarised in Table lV. The results in Table VI and VII, which are representative, attest to the presence of reduced regional cerebral flow in areas of the brain appropriate to the clinical findings and their change as these findings alter. Clinical studies in patients before and after cerebrovascular surgery are to be presented elsewhere.

The range of carotid flow characteristics as assessed by variance, skewness and kurtosis, Table V, show large variations. Some very small variance in spectra meet our hypothesis of increased edge velocity in some patients with turbulent flow due to small smooth projections. Some very wide spectra meet our hypothesis of eddying in turbulent flow due to sharp irregular projections affecting the central high velocity components. Clinical analysis of these results in relation to angiography and surgical findings form the subject of another study.

In summary we have described the technical and analytical problems of measuring cerebral flow noninvasively in millilitres per minute. Some clinical results are given to help validate the use of the method. We have discussed the theory and practice of deconvolution analysis applied to the flow characteristics of the carotid arteries in patients with carotid and cerebrovascular disease so that stenoses insufficient to cause a drop of volume flow rate may be studied. By the development of these techniques, it is hoped: to advance high technology preventative medicine through the objective assessment of patients at risk from stroke; to help in deciding when there is a need for angiography and intervention; and to assess the progress of patients before and after vascular surgery.

Acknowledgements

We are grateful to Professor Taylor and Mr. John Lumley for encouraging us to study their patients. We thank the St. Bartholomew's Hospital Joint Research Board and the Wellcome Trust for financial support for M.G..

REFERENCES

1. J. G. Brice, D.T. Dowsett, and R.D. Lowe "Haemodynamic effects of carotid artery stenosis" Brit. Med. J. 2, 1363 (1964).

2. K.E. Britton, G. Duplock and G.V. Hudson, "Carotid flow disorders in patients with cerebrovascular disease: their atraumatic assessment by carotigraphy" Brit. J. Surg. 61, 873 (1974).

3. D.H. Ingvar, S. Cronquist, R. Ekjerb, J. Risberg, and K. Hoedt-Rasmussen. "Normal values of cerebral blood flow in man including flow and weight estimates

of grey and white matter"
Acta neurol. scand. Suppl. 14, 39 (1965).

4. B.L. Mallett, and N. Veall, "Measurement of regional cerebral clearance rates in man using xenon-133 inhalation and extra cranial recording"
Clin. Sci. 29, 179 (1965).

5. G. Austin, D. Laffin, S. Rouhe, M.R. Edwards, and W. Hayward. "An non invasive intravenous isotope method for measuring cerebral blood flow"
Stroke, 4, 362 (1973).

6. K.E. Britton, C.C. Nimmon, P.H. Jarritt, M. Granowska, T.Y. Lee and J.M. McAlister, "Cerebrovascular disorder; assessment with radionuclides" in
Advanced Medicine, 13, G.M. Besser, Editor.
Pitman Medical, London, in press, 1977.

7. W.H. Olendorf, "Molecular criteria for blood-brain barrier penetration"
p. 17 in Non invasive brain imaging H.J. DeBlanc and J.A. Sorenson, Editors, Society of Nuclear Medicine, New York, 1976.

8. K.E. Britton, "A feasibility study of a screening test for abnormalities of carotid blood flow using technetium 99-m
J. Neurol. Neurosurg. Psychiat. 33, 379 (1969).

9. R.J. Cowan, C.D. Maynard, I. Meschan, R. Janeway, and K. Shigeno, "Value of routine use of the cerebral dynamic radioisotope study"
Radiol. 107, 111 (1973).

10. D.C. Moses, A.E. James, H.W. Strauss, and H.N. Wagner "Regional cerebral blood flow estimation in the diagnosis of cerebrovascular disease"
J. Nucl. Med. 13, 135 (1972).

11. D.C. Moses, T.K. Natarajan, T.J. Previosi, G.B. Udvarhelyi, and H.N. Wagner, "Quantitative cerebral circulation studies with sodium pertechnetate".
J. Nucl. Med. 14, 142 (1973).

12. F.H. Deland "Clinical application of cerebral dynamic perfusion studies"
p. 79 in Dynamic studies with radioisotopes in medicine, vol 2, International Atomic Energy Agency, Vienna, 1974.

13. A.H. Deutchman, L.R. Fulmer, G.R. Wise, D.W. Miller, and W.W. Hunter, "A clinically practical procedure for the quantitative evaluation of regional cerebral circulation"
J. Nucl. Med. 16, 525 (1975).

14. T.F. Budinger, F.H. Deland, H.E. Duggan, J.J. Bouz, B. Hoop, W.T. McLaughlin and P.M. Weber.
"Dynamic time-dependent analysis for computer assisted CNS studies"

p. 45, in Non Invasive brain imaging, H.J. Deblanc and J.A. Sorensen, Editors, Society of Nuclear Medicine New York.

15. H.N. Wagner, W.C. Klingensmith, L.G. Knowles, M.G. Lotter, and T.K. Natarajan "Advances in computer applications in radioactive tracer studies" p. 291 in Medical Radionuclide Imaging, Vol 2, International Atomic Energy Agency, Vienna, 1977.

16. J. Goddard, I.M.S. Wilkinson, H.J. Testa, and B.R. Pullan "Atraumatic cerebral blood flow estimation in men" p. 158 in Clinical blood flow measurement, J. Woodcock, Editor, Sector Publishing, London, 1976.

17. K.E. Britton, N.J.G. Brown, F. Cruz, H.C. Chang, D. Ralphs, and M. Myers, "Deconvolution analysis of dynamic studies" p 244 in Proc IVth Int. Conf on Information processing in Scintigraphy, C. Raynaud and A. Todd-Pokropek, Editors, Atomic Energy Commission, Orsay, Paris, 1975.

18. K.L. Zierler, "Theoretical basis of indicator dilution methods for measuring flow and volume" Circ. Res. 10, 393 (1962).

19. G. Taylor, "Dispersion of soluble matter in solvent flowing slowly through a tube" Proc. Royal Soc. A. 219, 186 (1953).

20. D. Lane, and J.A. Sirs, "Indicator dilution measurement of mean transit time and flow in a straight tube" J. Physics, E., 7, 51 (1974).

21. C.G. Caro, "The dispersion of indicator flowing through simplified models of the circulation and its relevance to velocity profile in blood vessels" J. Physiol. 185, 501 (1966).

22. J. Gamel, W.F., Rousseau, C.R. Katholo, and E. Mesel "Pitfalls in digital computation of the impulse response of vascular beds from indicator-dilution curves" Circ. Res. 32, 516 (1973).

23. M. Buhl, P. Charles, and F.T. Jensen "Elimination of scalp blood flow by head band during dynamic brain scintigraphy" J. Nucl. Med., 16, 679 (1975).

24. J. Marshall, "Cerebral blood flow" in The management of cerebrovascular disease 3rd edition, Blackwell, Oxford, 1976.

25. E. Zilkha, G. Ladurnar, L.D. Iliff, G.H. Du Boulay and J. Marshall, "Computer subtraction in regional cerebral blood volume measurements using the EMI-scanner" Brit. J. Radiol. 49, 330, (1976).

DISCUSSION:

Altschuler asked Britton if he had studied the time course of the development of cerebral circulatory compromise and, further, if he could determine whether such compromise is related to large or small vessel disease.

Britton responded by stating that they anticipate future studies to answer important clinical questions, but to be certain of their results they have limited their current effort to refining these methods. Dr. Britton stated that over the next two years they will have evaluated the method from clinical, mathematical, and modeling aspects and will then be able to study circulatory pathophysiology and the effects of therapeutic intervention.

Dr. Britton noted that the probe system is used to evaluate large vessel stenosis, but one should be able to detect small vessel disease from the cerebral blood flow data obtained from the gamma camera.

ANALYSIS OF ERRORS AND SYSTEMATIC BIASSES IN THE CALCULATION OF THE RENAL RETENTION FUNCTION.

F. Erbsmann°, J. Struyven[x], H. Ham°, A. Piepsz*.
° Department of Radioisotopes, St-Peter's Hospital,
x Department of Radiology, Brugmann Hospital,
* Pediatric Clinic, St-Peter's Hospital,
Free University of Brussels, Brussels, Belgium.

ABSTRACT

The renal retention function (RRF) of a given radio-tracer is defined as the time activity curve measured over one kidney after injection of a bolus of tracer in the renal artery. In practice the RRF can be computed by deconvolution of a kidney curve by a selected input curve. This paper deals with the effect of various possible errors in the calculation of the RRF. Data from routine patient files, simulation and direct renal artery injection are analyzed. It is found that the calculated RRF fits in acceptable limits to direct measurement. Deconvolution is an auto-corrective process. The RRF is little affected by the level of non renal activity correction and the choice of input curve (precordial or hepatic). Furthermore noise in the input curve has little influence on the RRF; preprocessing the input curve is useless and can generate biasses.

INTRODUCTION

The renal retention function (RRF) of a radio-tracer is defined as the time activity-curve measured over one kidney after injection of the tracer in the renal artery. This function is also called the renal impulse response function.

If the tracer is injected intravenously there is an exact mathematical relation between the activity in the kidney (R), the plasmatic concentration of the tracer (P) and the RRF (G)

$$R = G * P \tag{1}$$

where * stands for convolution operation.

Equation (1) is valid for linear and stationary systems. Cases for which the kidney is not stationary (eq. shows irregular pelvic contractions) were discarded from this study.

In kidney studies equation (1) can be rewritten as

$$G = R * P^{-1}. \tag{2}$$

P^{-1} is not an analytical function. In the case of discrete data however a set of numbers can be computed which give a good approximation of P^{-1}.

Several authors[1, 2, 3] have calculated the RRF by deconvoluting a kidney curve with a selected input curve representative of blood concentration of the tracer.

Uncertainties remain concerning :

a) error amplitude and error propagation due to noise in the renogram and the blood curve.

b) systematic biasses due to the choice of the input curve and the amount of background subtraction, or introduced by the processing of the data.

This paper deals with the effect of various possible errors in the calculation of the RRF.

Three types of data were studied

a) patient data for the analysis of background subtraction and choice of input curve.

b) simulated data for the study of error amplitude and error propagation.

c) direct intra-arterial injection was performed in a limited number of cases in order to make a comparison between measured and calculated RRF.

MATERIAL AND METHODS

Renal studies are performed routinely in our laboratory for the calculation of the RRF and the separate glomerular filtration rate[4].

The patient previously hydrated lies in prone position. A dosis of 5 mCu of 99mTc-DTPA is administered. Data are acquired with a Pho-gamma 3 scintillation camera linked with a Simis 3 computer. A low energy diverging collimator is used in order to obtain regions of interest over both kidneys and the heart. The variation in sensitivity accross the field of the collimated detector was measured and found to be negligible[5]. Data are sampled at 20 seconds time interval. Background was measured in a perirenal area of one or two channels width.

a) Patient data were selected from routine files. Eight kidneys were chosen with separate glomerular filtration rate ranging from 60 ml/min. to 10 ml/min. and maximal transit times ranging from 4 to over 20 minutes. Background was subtracted from the renogram in levels ranging from 0 to 2 and the kidney curves thus obtained were deconvoluted by using the matrix algorithm[3], the input function was measured either over the heart or the liver.

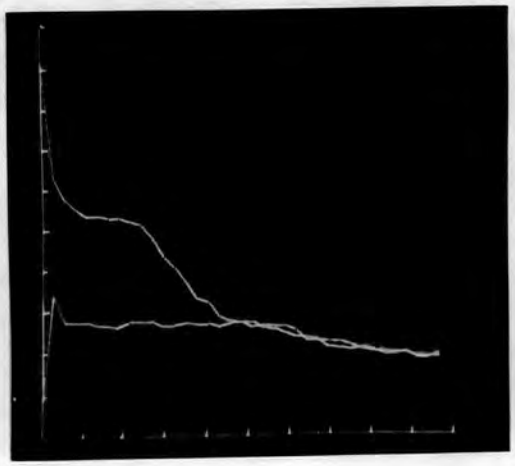
b) Simulated data were obtained by fitting smooth curves on the RRF, the blood curve, and background curve of the patient data. The renogram was reconstructed from these data. Poisson-noise was added on the blood curve, the background and the renogram. Deconvolution was then performed for the study of error amplitude and propagation.

c) Direct renal artery injection of 1 mCu radio-tracer was performed on 6 patients subjected to contrast renal arteriography. A routine renogram (see a) was obtained on those patients, less than 24 hours prior to the arteriography. Hydration conditions were identical. Only minute amounts of diluted contrast material were injected, prior to direct renal artery injection.

A collimated 2 inch sodium-iodide crystal was used, connected with a multi-channel analyzer. A lead shielding was positionned under radioscopic control in order to ensure ideal collimation of the kidney. Data are sampled at 4 sec. time intervals.

A correction had to be performed on the raw data due to the recirculation of approximatively 80 % of the injected isotope. It was assumed that this portion behaves like an IV injection, so the renogram performed the day before can be subtracted from the raw data after proper scaling (Fig. 1).

Comparison between measured and computed curves was performed by computing the mean transit times (TM) by two methods : one where only positive values of the RRF were used (TM1), the other where all the available data were used (TM2); other values like T_{80}, T_{50}, T_{20} were also computed; they correspond to the moment where 80 %, 50 %, 20 % of the activity remains in the kidney.



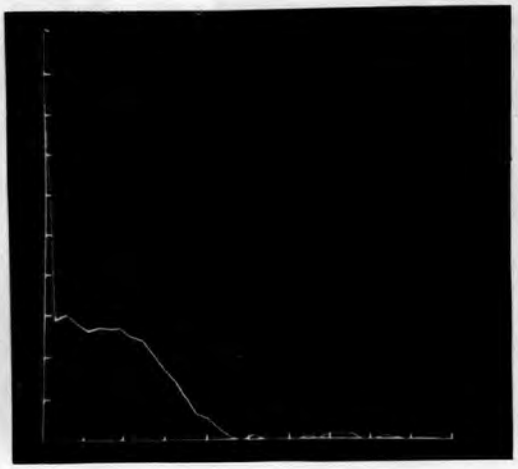


Fig. 1a. Raw data from direct renal artery injection and corresponding renogram properly scaled.

Fig. 1b. Result of subtracting renogram from raw data. Initial peak is due to recirculated activity. (sampling time 20 seconds, scale on time axis : 1 minute).

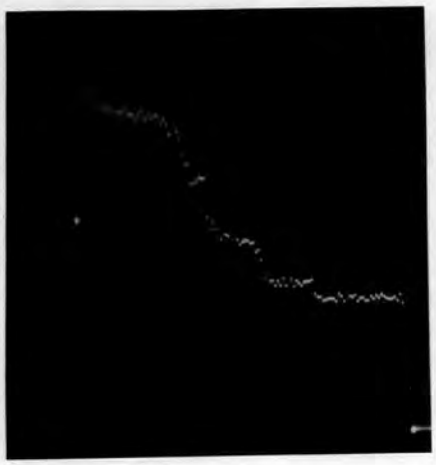


Fig. 2. Raw data from artery injection in fasting subject. (sampling time 4 seconds).

The importance of proper hydration can be seen in Fig. 2, showing the raw data obtained in a fasting subject. After the plateau, the activity within the kidney decreases in steps spaced by about one minute. On hydrated patients these steps are at the limit of detectability and spaced by about 20 seconds.

A possible interpretation of these steps is pelvic contractions which produce a sharp ejection of the activity out of the field of the detector.

RESULTS

1. Error amplitude

Exact calculation of the RRF was performed on simulated data after addition of pseudo-random noise. The process was repeated ten times. The results are shown superimposed in Fig. 3 a, b, c.

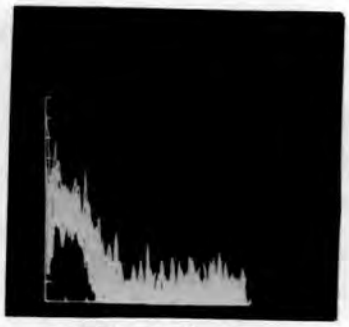


Fig. 3 a
Noise on all data

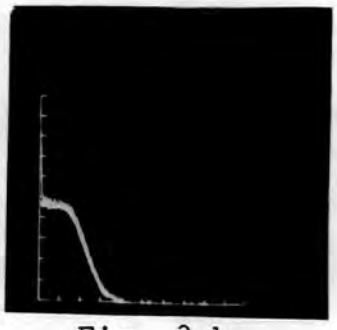


Fig. 3 b
Noise on input curve

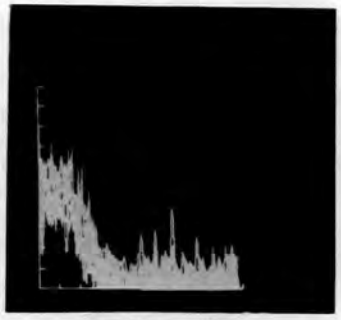


Fig. 3 c
Noise on renogram

These results indicate that the error amplitude in RRF is constant (Fig. 3 a). There is no long term build-up of errors. Furthermore, noise in the input curve only produce errors of small amplitude in the RRF (Fig. 3 b). Most of the information on the RRF is carried by the renogram, the RRF will thus be more sensitive to little variations in the renogram (Fig. 3 c).

2. Noise propagation.

Noise propagation was studied by adding a single small error in one data-point. This small error E is assimilable to a delta function.

2a. Errors appearing in the kidney curve.--Errors appearing in the kidney curve will affect the RRF in a linear way. This effect is independent of the moment of occurence (Stationarity)

$$G' = (R + E) * P^{-1} = R * P^{-1} + E * P^{-1} \quad (3)$$

$$G' = G + e \times P^{-1} \quad (4)$$

were G' indicate that an error has been added.

Table 1. Numerical values of P^{-1}

Time (min.sec.)	P^{-1}
.20	100
.40	- 61
1.00	- 12
1.20	- 2.8
1.40	- 1.2
2.00	- 0.8

The propagation pattern of E in G is thus proportional to P^{-1}. Numerical values of P^{-1} are listed in table I. The values of P^{-1} oscillate only in the first 2 points and decrease quickly in absolute value, becoming negligible after 2 minutes.

2b. Errors in the blood curve.--Errors appearing in the blood curve affect P^{-1} in a non stationary and non linear way. The derivation of first order approximations is cumbersome.

Error propagation was evaluated by adding a 1 % error in selected data points of P. Then P'^{-1} is computed and compared with P^{-1}, the relative errors are shown in table 2.

Table 2. Propagation of errors on P^{-1}

I \ K	0	1	2	3	4	5
1	1.	- 2.	1.	3.	2.1	.6
5	28.	- 50.	7.	5.	1.4	.9
10	38.	- 54.	7.	5.	1.7	.4
15	48.	- 68.	9.	6.	2.1	.5
20	66.	- 93.	13.	9.	2.9	.7
25	93.	-132.	18.	13.	4.3	1.0

A 1 % error was added in the Ith data point of P. This generated errors in the point I, I + 1, I + 2, ..., I + K, ... of P^{-1}. Table shows relative errors of P^{-1} for various values of I and K (sampling rate 20 seconds).

The values of the errors are important only in the first points and then quickly vanish to naught.

The important relative values are in fact of little absolute amplitude (compare table 1 and 2). Furthermore most of the oscillation will be smoothed out by the convolution of P^{-1} with R (Formula 2). This holds only for small variations. In the case of important errors in P (biasses) the effect will be felt over the entire RRF[6].

3. Effect of non renal activity subtraction and choice of input curve.

Non renal activity subtraction acts linearly on RRF. Its effect can be predicted by computing the function

$$F = B * P^{-1} \qquad (5)$$

where B is the background activity.

The function F shows a sharp peak at the origin followed by a slow component of less than 5 % amplitude. This component is negligible after 10 minutes and can be interpreted as due to the transit of the marker in extra-vascular space.

Subtracting increasing amounts of background (Fig. 4) will reduce the height of the plateau slightly. The transit time distribution is not affected significantly.

The choice of the input curve, (precordial or hepatic) does not affect the results significantly (Fig. 5).

Fig. 4. Effect of background subtraction.

Fig. 5. RRF calculated with liver and heart input curve.

4. Comparison of computed RRF with direct renal artery injection.

Two examples of comparison are shown in Fig. 6. The parameter T_{80} was found to be highly unreliable due to difficulties in establishing exactly the height of the plateau in computed data.

T_{50} and T_{20} behave in a way very similar to TM1 and TM2. Individual values of TM2 are shown in table 3.

Table 4 indicate the mean difference between true values and TM1 and TM2 for RRF's calculated with hepatic or precordial input functions.

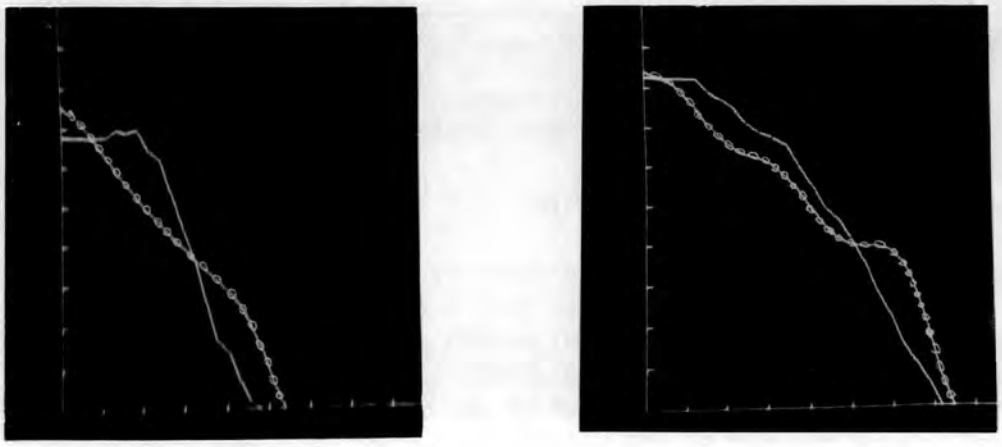


Fig. 6. Comparison of calculated RRF (o-o-o) and measured RRF (solid line) in two cases. (scale on time axis 1 minute).

Table 3. Comparison between calculated and computed transit times. (TM2)

Patient number	Intra-arterial	Hepatic input curve	Precordial input curve
1	234	212	290
2	314	290	344
3	256	226	236
4	232	257	274
5	334	278	332
6	244	316	362

(Time in seconds)

Table 4.

	Hepatic input curve	Precordial input curve
TM1	15 s	51 s
TM2	- 5 s	37 s

Mean difference between calculated and measured mean transit times. The standard deviations were close to 40 s.

CONCLUSION

The mathematical procedures in the computation of the RRF are exact. They only differ in the way noise is handled, by smoothing, bounding or fitting exponentials.

Our results indicate that preprocessing the input curve is of no interest, mild smoothing does not modify noticeably the RRF, whereas sophisticated procedures like data bounding introduce biasses.

The amplitude of the error on the RRF is proportional to the error in the renogram. It was found in practice that 6 unweighted 3 point smoothing were sufficient to reduce statistical errors within acceptable limits. An exact algorithm for computing the error amplitude[7] is based on the use of the square matrix of P^{-1}.

No indication was found that background subtraction on the kidney curve would increase the reliability of the data, there is little difference between using hepatic and precordial curves.

The computed RRF fits well with data obtained from direct renal artery injection.

ACKNOWLEDGEMENTS

Our warmest thanks go to Prof. A.M. Ermans for his constant support to this work. We would also like to thank Dr. A. Schoutens, Prof. J. Dumont, Mr. M. Martins, Mr. A. Dobbeleir and Mr. J.L. Froideville, without whom this work would not have been possible. Dr. B. Diffey helped us with many fruitful discussions.

REFERENCES

1. N.J. Brown and K.E. Britton, "The Theory of Renography and Analysis of results," Radionuclides in Nephrology, M.D. Blaufox and J.L. Funck-Brentano, Editors, Grune and Stratton, 1972.

2. R.W. Kenny, D.M. Ackery, J.S. Fleming, B.A. Goddard, R.W. Grant, "Deconvolution analysis of scintillation camera renogram," Br. J. Radiol., A8, 481 (1975).

3. B.L. Diffey, F.M. Hall, J.R. Corfield, "The ^{99m}Tc-DTPA Dynamic, renal scan with deconvolution analysis," J. Nucl. Med., 17, 353 (1976).

4. A. Piepsz, A. Dobbeleir, F. Erbsmann, "Détermination de la Clearance Séparée des Reins à l'aide du complexe Tc^{99m}-DTPA et de la Caméra à Scintillation," Proc. of XVII Colloque de Médecine Nucléaire de la Langue Française, C. Kellersohn and C. Raynaud, Editors, (1975).

5. F. Erbsmann, J. Paternot, A. Piepsz, A. Dobbeleir, J.L. Froideville, "Sensibilité et Homogénéité," La Fixation Rénale du Mercure et du DMSA, C. Raynaud, Editor, CEA, (1975).

6. B.L. Diffey, Personal communication.

7. D.R. Dance, Personal communication.

DISCUSSION:

Brown stated that they have found that very often both kidneys will give similar renograms. If that were the case, during an arterial injection one could use the contralateral kidney to give an idea of the magnitude of error.

Erbsmann said that he agreed in principle, but unfortunately he did not have the equipment necessary to perform simultaneous studies on both kidneys.

Britton stated that the irregularities in the transit time spectra that were presented were probably due to fluctuations in input rather than output because when blood flow is reduced slightly the increase in salt and water absorption is exaggerated. Since a kidney with a normal pelvis has a six second residence time as a mean, it is unlikely that the fluctuations in the transit time distributions are related to pelvic contractions but are much more likely to be fluctuations in blood flow possibly related to the arterial catheter. Thus, a prolonged transit time is much easier to obtain by varying input by a small amount rather than varying output by a small amount. It is important also to recognize that the injection of contrast to place the angiogram catheter will cause an osmotic diuresis which will decrease the mean transit time.

Erbsmann stated that they inject the tracer shortly after the contrast agent and then immediately withdraw the catheter.

VALIDATION OF RENAL TRANSIT-TIME CALCULATIONS USING COMPARTMENTAL MODELS AND DIRECT MEASUREMENTS

Ronald R. Price, J.J. Touya, R. Branch,
J. Goddard and A.B. Brill
Vanderbilt University Medical Center
Nashville, Tennessee U.S.A.

ABSTRACT

This work presents an investigation into the accuracy and limitations of the deconvolution techniques used in our laboratory for estimating renal transit-time distributions. These investigations have utilized both analytical (compartmental modelling) methods and direct measurements in experimental animals. Compartmental modelling was used to study the effects of random noise, data smoothing and filtering, blood background, variations in data sampling intervals, non-linear systems and multiple transfer paths. Direct measurements have been used to establish the precision of the techniques through repeated measurements in dogs. The accuracy of the techniques have been tested over a wide range of renal function through the use of direct renal artery injections in dogs.

Serial studies extending over a period of 50 days in a patient with a newly transplanted kidney are presented. The sequential transit-time distributions are compared with the sequential renograms.

Supported in part by ERDA Contract EY-76-S-05-2401

VALIDATION OF RENAL TRANSIT-TIME CALCULATIONS USING COMPARTMENTAL MODELS AND DIRECT MEASUREMENTS

Ronald R. Price, J.J. Touya, R. Branch,
J. Goddard and A.B. Brill
Vanderbilt University Medical Center
Nashville, Tennessee U.S.A.

Introduction

Current interest in the clinical utility of the renal transit-time distribution of I-131 labelled Orthoiodohippurate (OIH) has prompted us to explore a number of mathematical deconvolution techniques and to investigate the inherent uncertainties associated with each. Our interest in the technique was generated in part by an earlier work of Drs. Brown and Britton (1). Their work indicated that the renal transit-time distribution of I-131-OIH contained information regarding regional renal blood flow. This information is particularly relevant to our ongoing study of renal vascular hypertension.

Although the literature contains a number of reports on the clinical applications of the deconvolution techniques (2-4), few studies designed to validate or test the limitations of the techniques have been presented. In this work we present our investigations into the accuracy and limitations of the deconvolution technique as used in our laboratory. To this end, our work has included the use of both analytical techniques (compartmental modelling) and direct measurements in prepared animal models.

Mathematical deconvolution of the renogram curve requires that the kidney be regarded as a linear system and that all of the system parameters remain constant in time. If these assumptions are satisfied then the output of the kidney will be a convolution of the input function (the measured blood level of the ^{131}I-OIH) and the transfer function of the kidney. It is the purpose of the deconvolution calculation to determine the transfer function (transit-time distribution). The solution to the deconvolution equation for the transfer function can be carried out using a number of well known techniques if the input and output curves are known. An equivalent calculation can be carried out using the composite renogram in place of the output curve. The resulting transfer function will be related to the integral of the transit-time distribution and will be referred to later as the retention function. We have used two deconvolution techniques: 1) Fourier Tranform and 2) a discrete point-by-point calculation. Both can be carried out easily and rapidly on a small clinical computer.

Compartmental Model

A schematic of the basic compartmental model used in these studies is shown in Figure 1. Even though this model has been used successfully by others (5) to describe renal function in a variety disease states, the physiological accuracy of the model is not of central importance to our work. The necessary

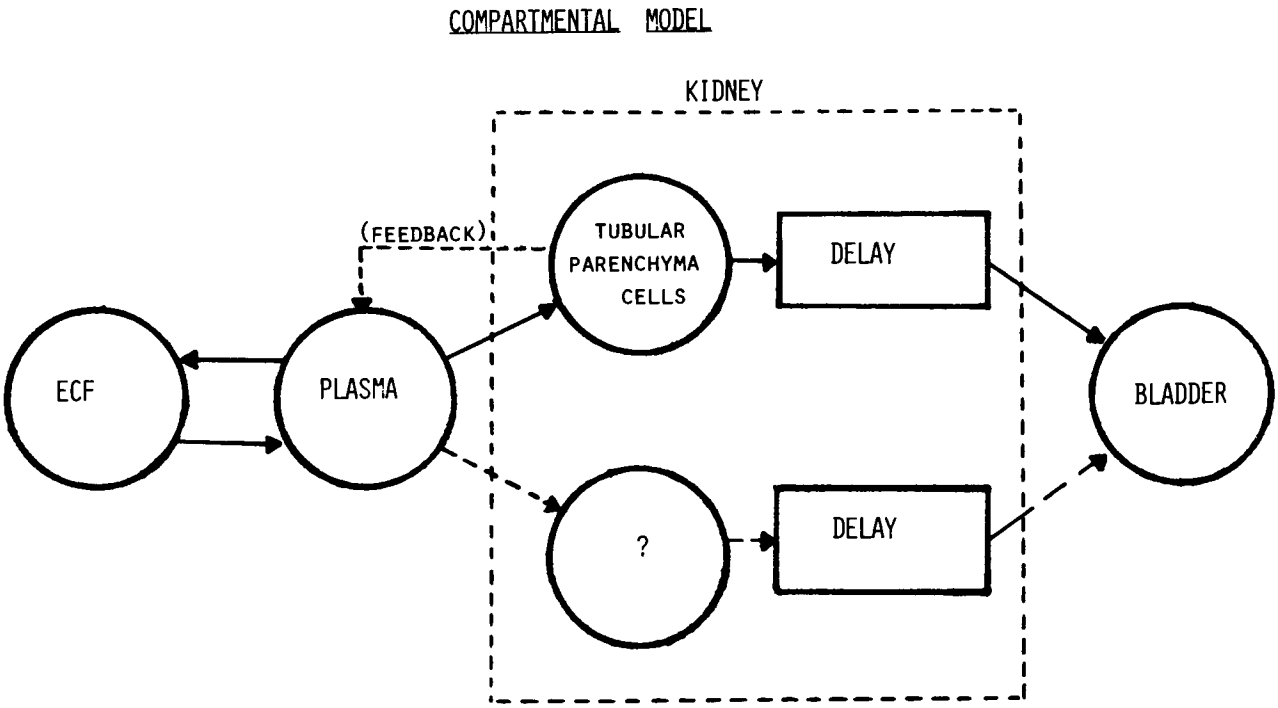


Figure 1. Schematic of the compartmental model used to simulate I-131-OIH kinetics

requirements of the model for our purposes are: (1) that it employ algorithms which are independent of the deconvolution calculations and (2) it have sufficient complexity and flexibility to yield a variety of realistic states of renal function.

We have found the model to be a flexible tool and have found it useful for investigating a variety of questions relating to the deconvolution calculations. These have included a study of the effects of:

1) random noise levels in the data

2) various data smoothing and filtering algorithms

3) blood background interference

4) variations in the data sampling intervals

5) non-linear systems (feedback), and

6) multiple paralled transfer paths.

The model includes plasma, extra-cellular fluid, "kidney" and bladder or urine compartments. The "kidney" is assumed to be composed of a single mixing compartment (labelled as "tubular cells" in the figure) plus a delaying component. To investigate the effects of multiple transfer paths within the kidney, we simply include additional equivalent pathways in parallel with the first (see figure). The simulated external measurement of the kidney in every case was taken as a linear combination of all the compartments and delays (dashed box). The intercompartmental rate constants were chosen to yield the transfer functions shown in Figure 2. The single pathway simulation presented in this work used a transfer function with a mean transit-time of 1.5 minutes. The curves presented to demonstrate two pathway systems used both of the transfer functions shown in Figure 2.

Of all the factors affecting the deconvolution calculations, we found for the I-131-OIH data that the errors introduced by random noise were the most significant. Simulated noise free kidney and blood activity curves are shown in Figure 3, along with the same curves after random noise has been added. The Poisson noise level depends upon the contents in each data interval ($\pm\sqrt{n}$). The maximum counts in the kidney curve in Figure 3 was 1,000 in a 10 second interval corresponding to an error of approximately 3% at the peak. In our institution using a scintillation camera, this count rate is obtained only with patients who have good renal function. From these curves the input and output curves were then generated. The input curve was taken to be proportional to the blood level and the output curve was determined by the technique of Britton and Brown where a kidney removal component is determined from the extrapolated kidney uptake. It is the derivative of the difference between the removal component and the measured renogram that yields the output curve.

Figure 4 shows the results of the output curve calculation and the corresponding transit-time distribution for the curves shown in Figure 3. The solid curve is the transit time curve corresponding to the expected distribution as determined from the noise free system. Even this relatively small amount of noise completely obscured even the major characteristics of the distribution. In an effort to resolve this problem, we have investigated a number of data filtering techniques. The results of applying one such filter is shown. This filter was applied only to the raw renogram data and not the blood curve and was designed to attenuate fluctuations in the data which had periods less than about one minute. The optimum filter must be chosen to suppress noise while still maintaining adequate time resolution. All of the data presented in this paper have been processed by this filter.

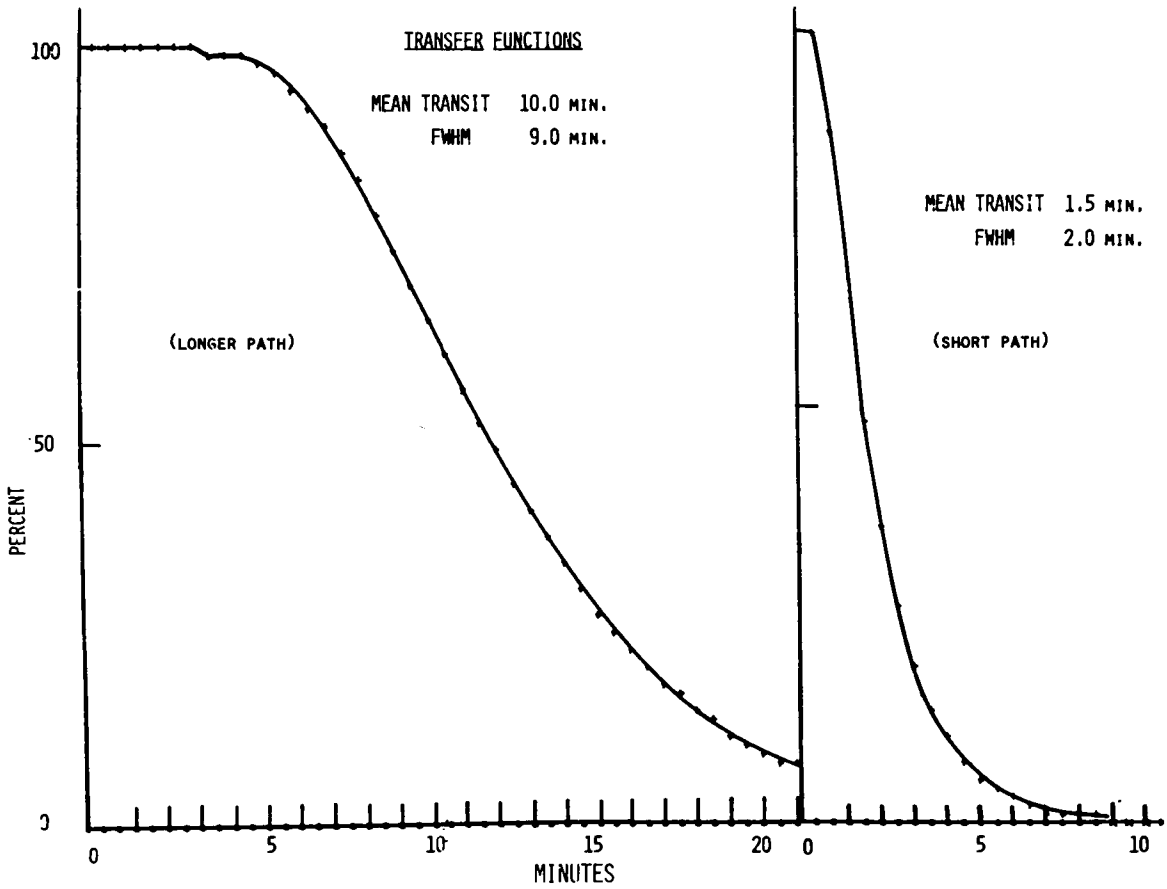
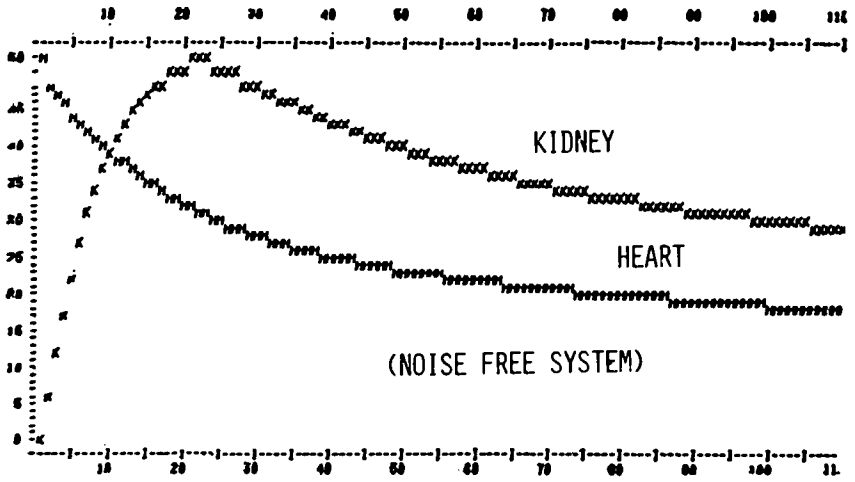


Figure 2. Renal transfer functions of I-131-OIH used to simulate renograms with single and multiple parallel pathways.

SIMULATION OF AN IDEAL SYSTEM



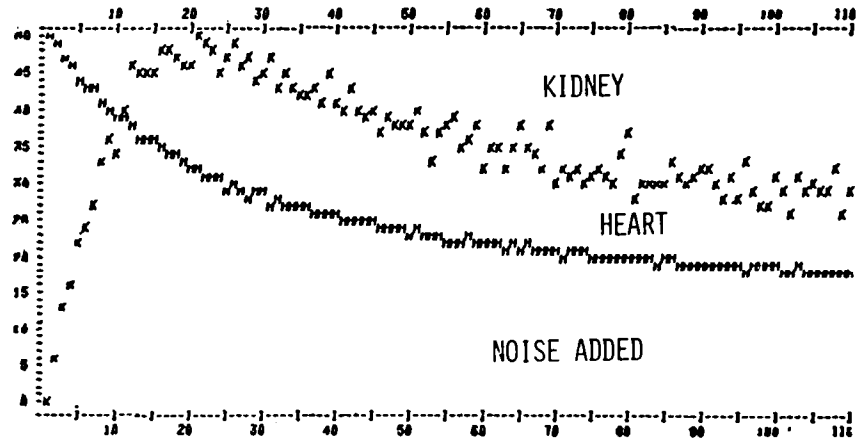


Figure 3. Simulated renogram and blood activity curve using a compartmental model. (above) Noise free simulation, (below) same curves after the addition of random noise. Maximum count in each curve is 1000.

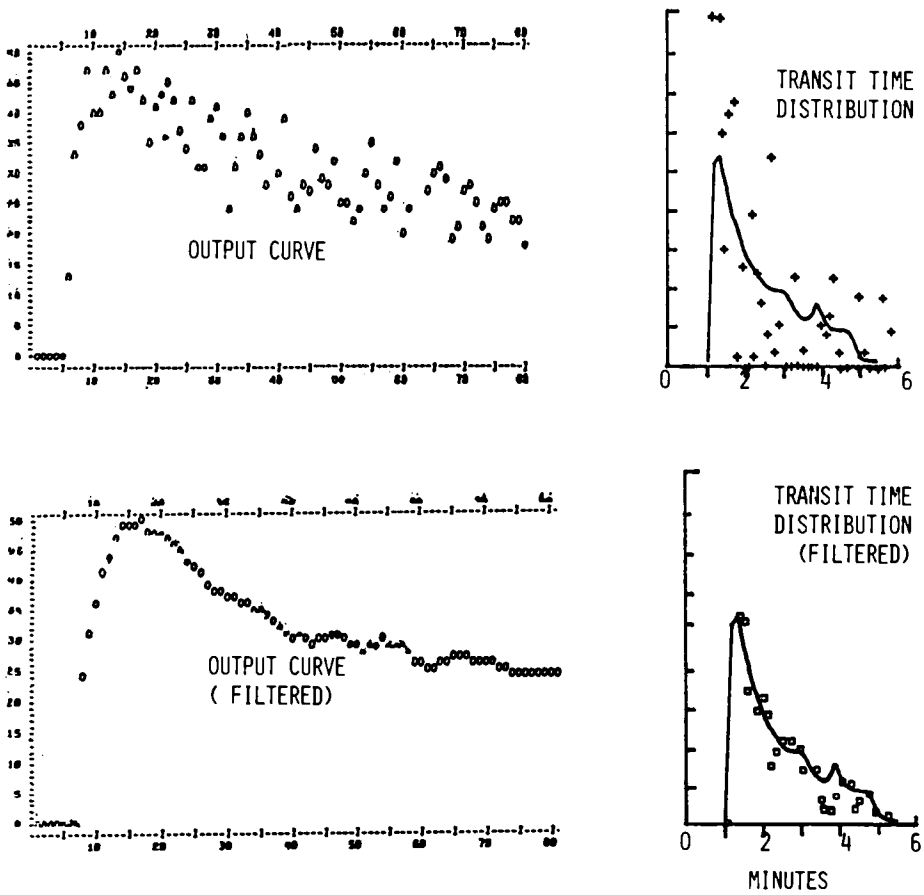


Figure 4. Renal output curves derived from renogram shown in Figure 3, (right) corresponding transit-time distribution. Solid line is the transit-time distribution for the noise free system. (Below) Same curves after processing with a low pass filter.

As noted in the introduction, one of the reasons for beginning this work was to look for a bimodal behavior in the transit-time spectrum. Britton and Brown have observed this structure in normal kidneys and have suggested that this curve structure reflects contributions from two distinct nephron populations. They also suggested that the relative areas under the two transit-time populations yield information on regional renal function. In order to assure ourselves that our bounding and filtering routines did not distort or filter out a bimodal structure in the transit time spectrum, we again utilized model simulated data.

In Figure 5 are shown simulated kidney and blood curves in which two distinct and parallel renal transfer functions are operating. The transfer functions illustrated in Figure 2 were used for these curves. Since the renogram is a convolution of the transfer function with the blood curve, the renogram can contain no frequencies higher than the narrowest transfer function. The criterion which the filter must satisfy, therefore, is whether or not the shortest transfer function will be preserved. Also included in Figure 5 are two different noise level simulations. The count rates were chosen to simulate typical data rates that might be observed in a Tc-99m DTPA renogram (5000,2500) and a typical I-131-OIH renogram (1000,500). The heart maximum was chosen to be twice the peak kidney count to approximate what is generally obtained with our measurement systems.

The same filter described earlier was used to process the curves in Figure 5. The results of the deconvolution calculations are shown in Figure 6. The solid line corresponds to the point-by-point calculation and the F's correspond to the Fourier transform calculation. The noise-free calculation (within round-off errors) was found to be identical to the expected analytical expressions formed from the sum of the two transfer functions. In comparison to the noise free curve, the higher count rate data (5000,2500) was found to have a 25% improvement in the Chi-square value over the lower count rate curve (1000,500). Especially in the point-by-point calculation the data processing introduced an uncertainty in the minimum transit-time. However, in every case the peak position and widths of the two components remained relatively stable. In the lower count rate data, additional (artifactual) components were beginning to appear, indicating a stronger filter may be warranted. Our decision was to stay with these filter parameters rather than introduce further distortions in the minimum transit-time.

The effects of varing levels of blood background is illustrated in Figure 7. The only effect on the retention function (shown by the "R" curves), was the spike artifact produced in the first element. This results when the blood background is high, since the kidney curve then tends to follow the blood curve exactly. By discarding the first element of the retention curve one may retrieve the correct retention function. When looking at the differential curve (transit-time distribution) we see that this will introduce an uncertainty in the minimum transit-time. The magnitude of the uncertainty depends upon the derivative interval.

The deconvolution-calculation assumes no reflex of the tracer from the kidney back into the circulating blood pool. In the normal, reflux is known to be minimal; however, for some radiopharmaceuticals and in some disease states reflux or the reappearance of the unbound nuclide may become sizeable. Compartmental model simulations were also used to study the distortion imposed upon the deconvolved renal transfer function as a function of the magnitude of refluxed or reappearing tracer.

Referring back to the model schematic (Fig. 1) reflux is easily introduced into the model by including a flow path from the kidney back to the plasma. Fairly large values for the feedback (25-50 percent of the forward flow) manifested themselves by very subtle changes in the renogram. As might be expected,

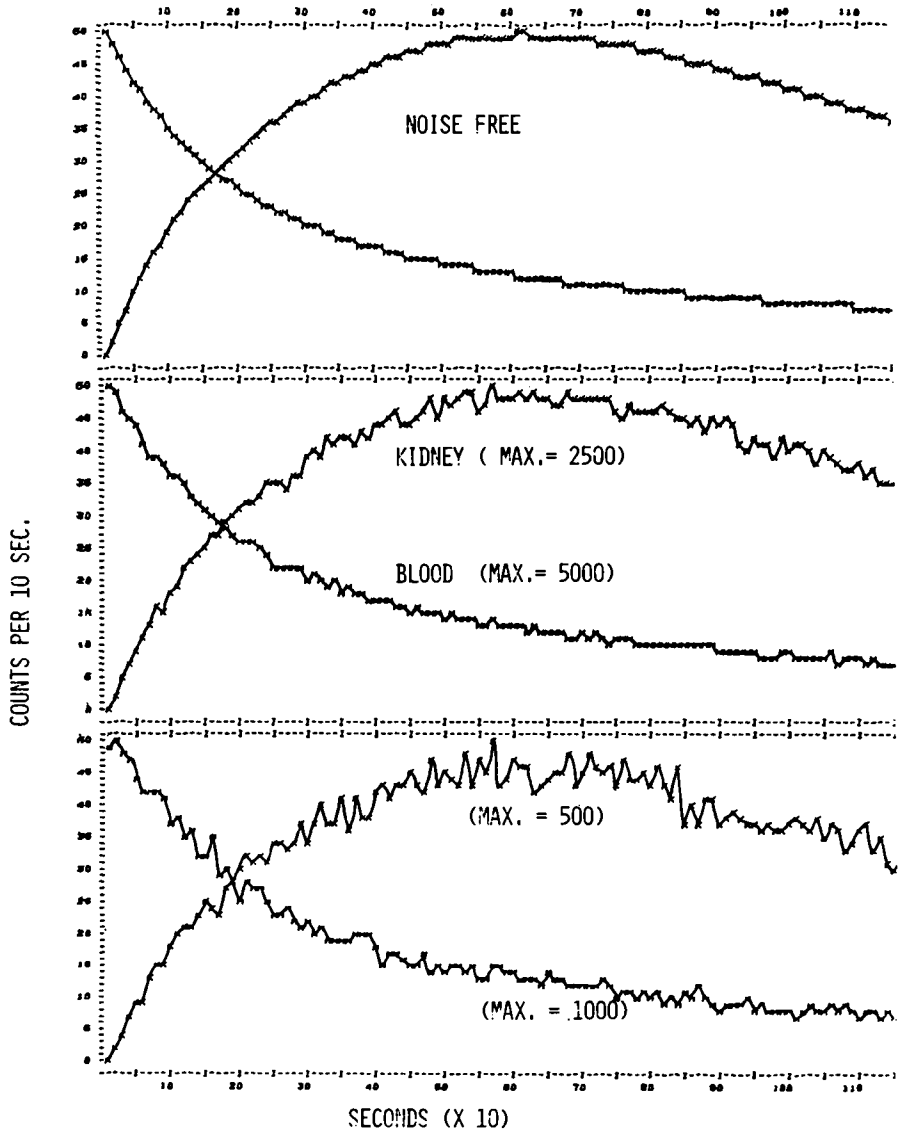


Figure 5. Simulated kidney and blood curves in which two transfer functions are operating. (Top) noise free, (middle) random noise added-kidney maximum count = 2500, blood maximum = 5000, (bottom) random noise added-kidney = 500, blood = 1000.

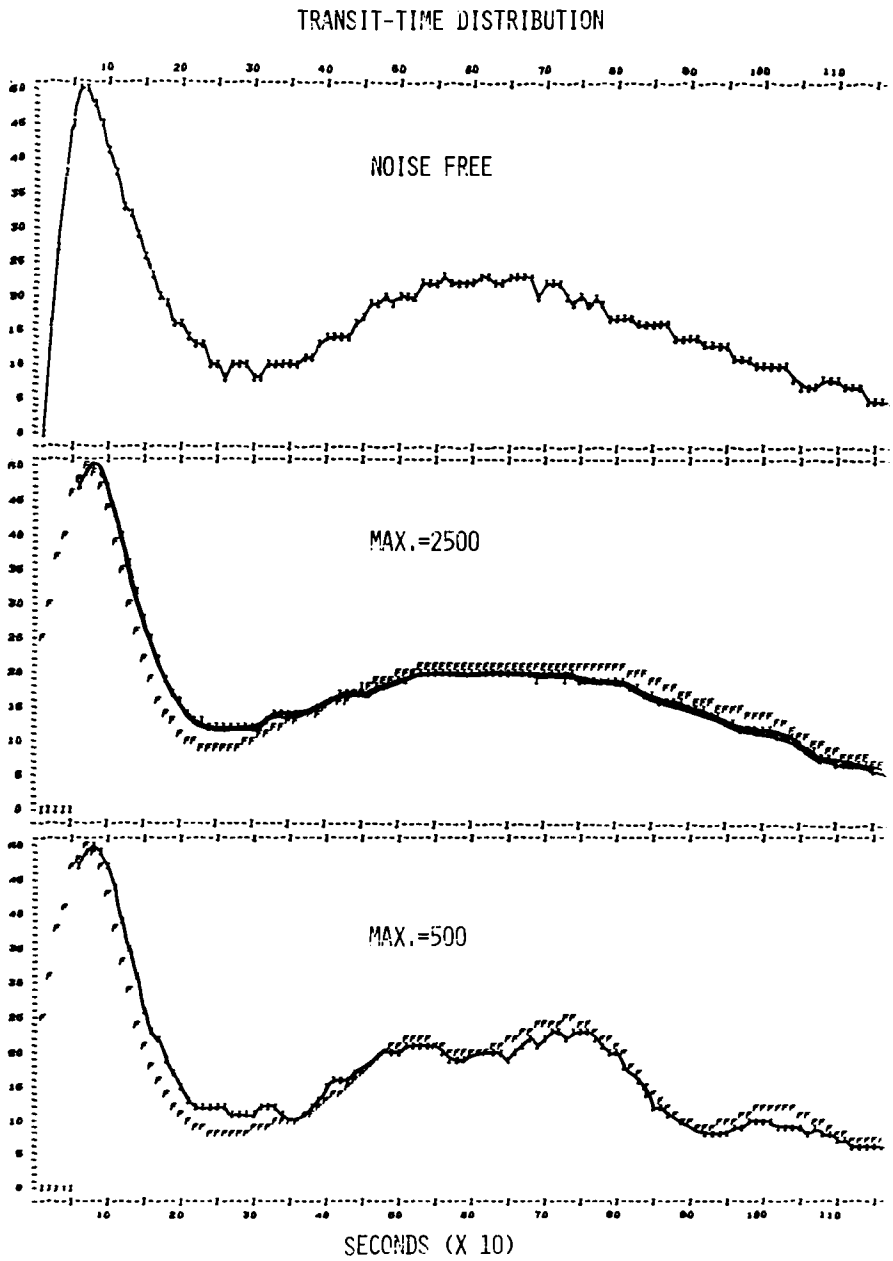


Figure 6. Deconvolved transit-time distributions corresponding to Figure 5. Solid line corresponds to the point-by-point discrete calculations and the F's correspond to the Fourier transform calculation.

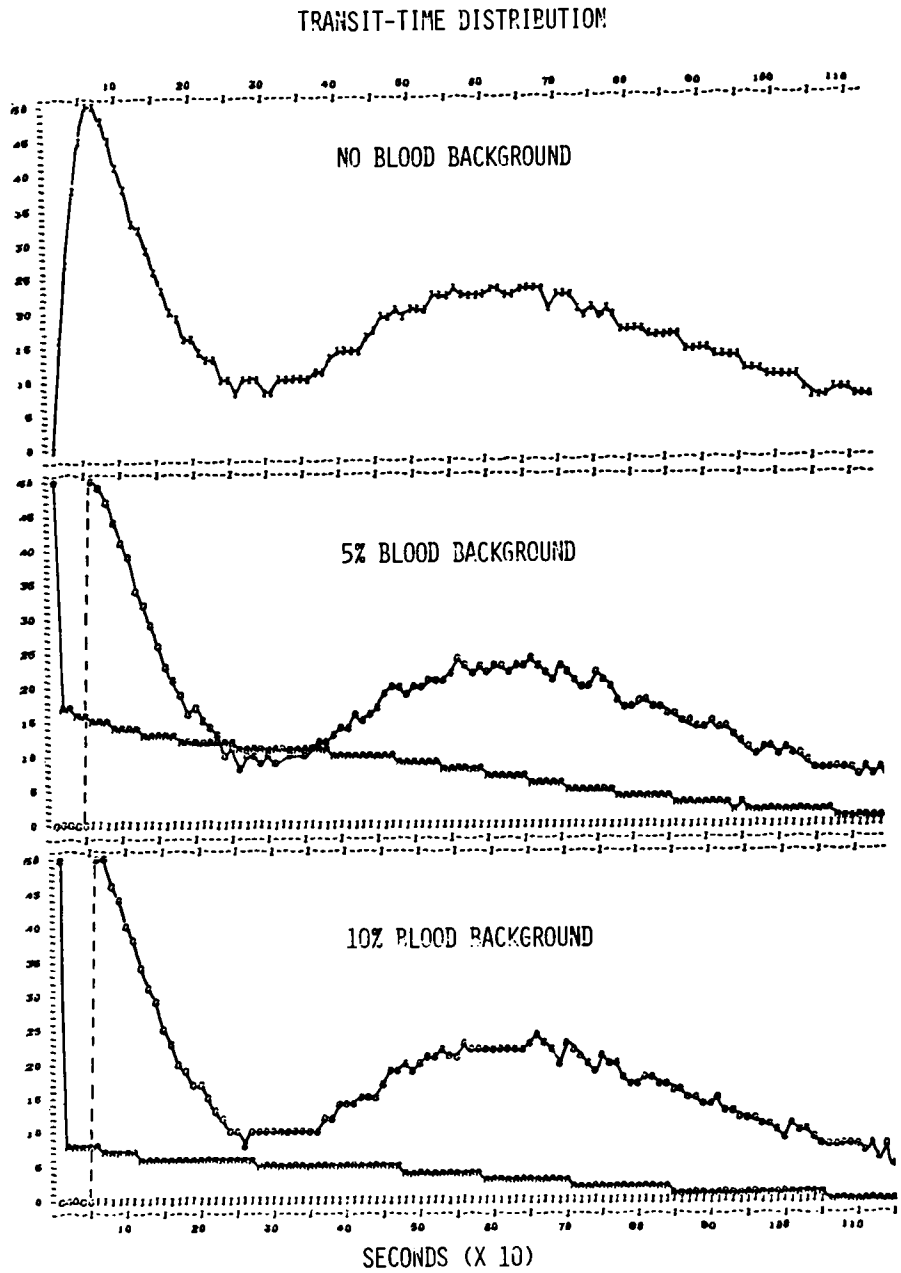


Figure 7. Deconvolved transit-time distributions (G's) as a function blood background. (Top) No blood background, (middle) blood background equal to 5% of the blood pool activity and (bottom) background equal to 10%. R's refer to retention functions.

the deconvolution calculation showed the new systems to have increasingly rapid transit-times as the magnitude of the feedback component increased. In Figure 8 are shown the retention functions corresponding to normal flow patterns compared with a system in which the feedback component was equal to 50% of the forward flow. The mean transit-time decreased from 90 seconds to 45 seconds in this case.

Direct Measurements

Having established the adequacy of the deconvolution calculations in simulated data we extended our work to controlled measurements in prepared dog models. Our first measurements were designed to test the precision of the techniques by making three sets of sequential measurements on the same dog maintained in the same functional state. Figure 9 shows the results of one of these studies. We concluded that the results were in good agreement.

The second phase of the validation was to test the accuracy of the technique by comparison with direct renal artery injections. Figure 10 shows two renograms taken from a dog in two distinctly different functional states. The decreased function state was created by the injection of drugs. The calculated transit-time distributions for each case are shown to the right. Note the difference in normalization and also the relative excretion percentages for the two cases. At the same time that these renograms (resulting from a peripheral IV injection) were made we measured the transfer function directly by making a bolus injection of I-131-OIH directly into the renal artery. Figure 11 shows the results of the direct measurement of the retention function as compared with the retention function derived from the deconvolution of the I.V. renogram. In both cases we see reasonably good agreement. In the case of the control study we see an approximately 90% extraction efficiency for OIH. This results in about 10% of the OIH passing through the kidney during the first pass and then becoming diluted over the blood pool. Thus 10% of the bolus looks very similiar to an I.V. injection. If one subtracts this component we get a result that almost exactly duplicates the deconvolution curve and is shown by the x's in the figure.

To check the program's calculation of the output curve a catheter was placed into the dog's ureter and urine samples were taken every 30 seconds. Within the counting errors of the urine samples, the urine activity curve compared favorably with the derived output curve.

From the results of the dog studies, in particular the intraarterial injection which has very good counting statistics, we have not seen any evidence of a bimodal character in these curves. If the effect is there it is very likely that the counting statistics available from an intravenous injection will not be adequate to resolve the second component. Rather the kidney seems to behave very similiar to a classical one-compartment system plus a delay.

Analysis of Transplant Renogram

After the validation phase we extended our work to study the transplanted kidney. Figures 12 and 13 show eight selected studies from a series of approximately 15 serial studies of a patient who had received a transplanted kidney. This particular patient was chosen because of the extreme variations in renal function ranging from excellent to poor function, including a rejection phase followed by a complete recovery to good function again. The retention function and the renogram plus bladder curve are presented for comparison. One day post operatively the patient had a BUN of 49 and a creatine clearance of 2.6. This continued to improve until day 12 (BUN = 26, creatine = 1.0) at which time he was discharged. Approximately one week later the patient returned running a temperature, urine volume had dropped to one-fourth (BUN = 200 and creatine =

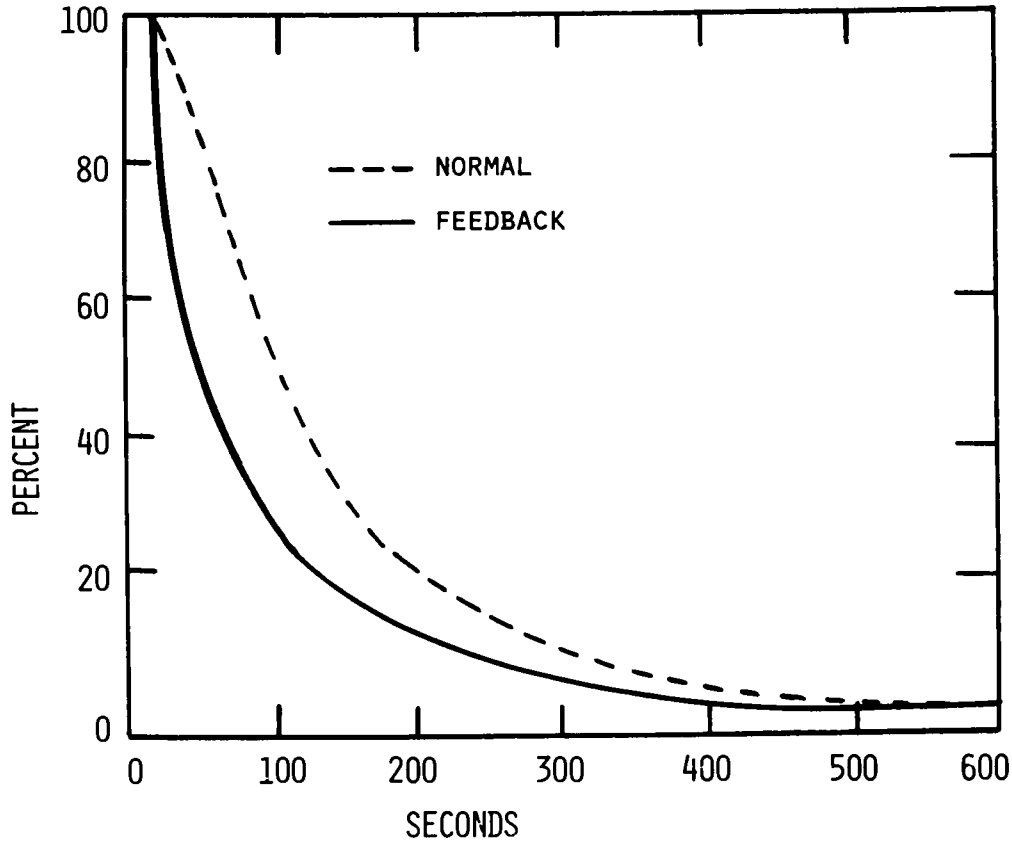


Figure 8. Retention function corresponding to a normal flow pattern compared to retention function of a system in which a feedback component equal to 50% of the forward flow was operating.

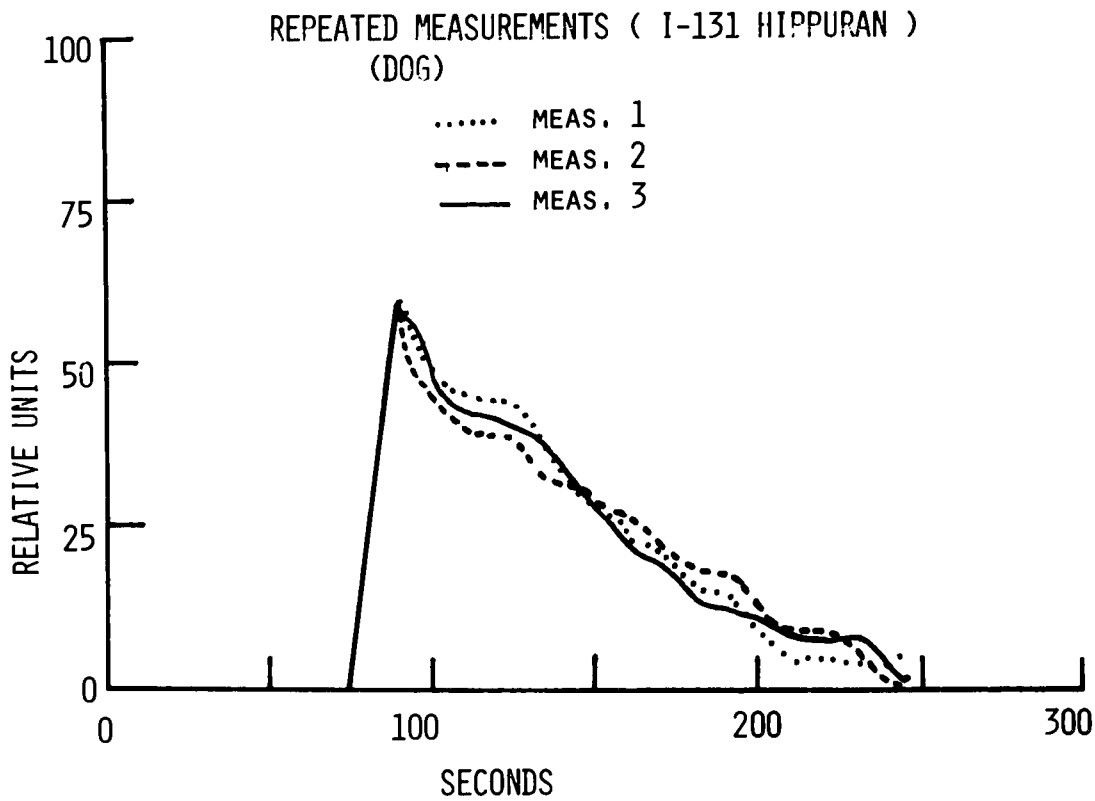


Figure 9. Tests of the precision of the deconvolved techniques. Transit-time distributions from sequential studies on the same dog maintained in the same functional state.

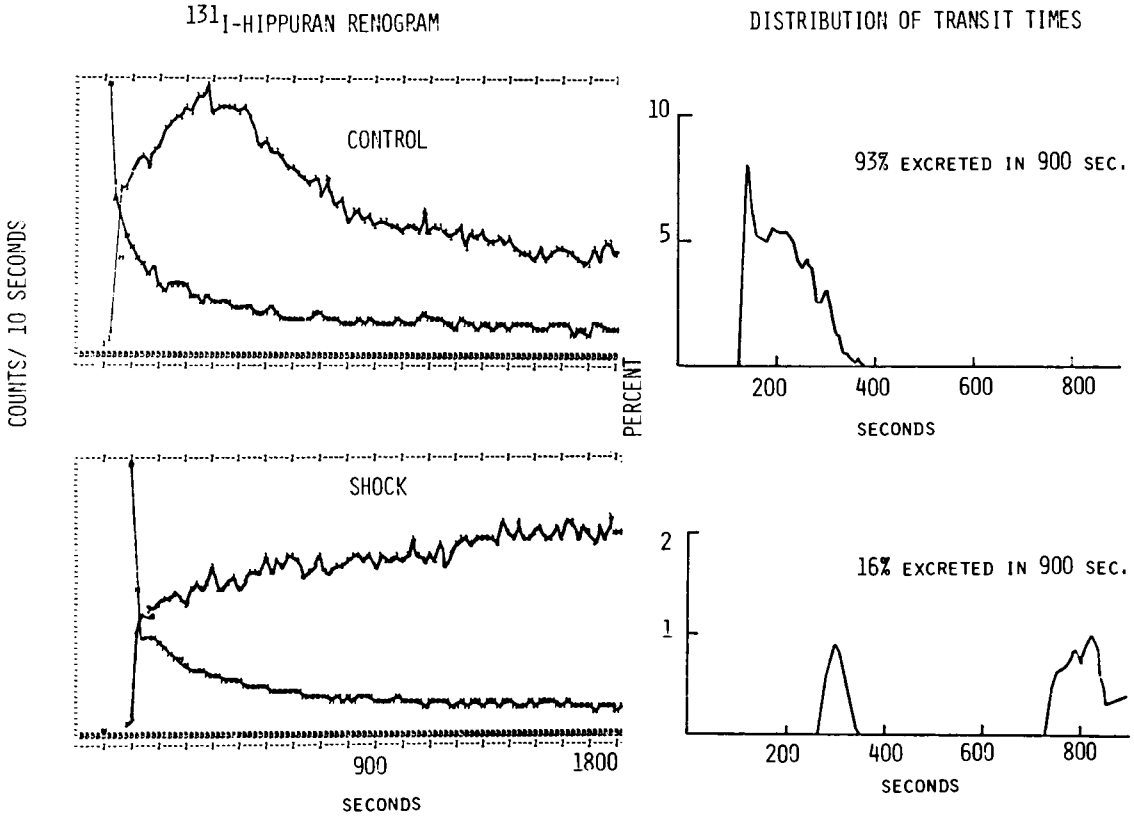


Figure 10. Renograms from a prepared dog in two different functional states. (Right) Corresponding deconvolved transit-time distributions.

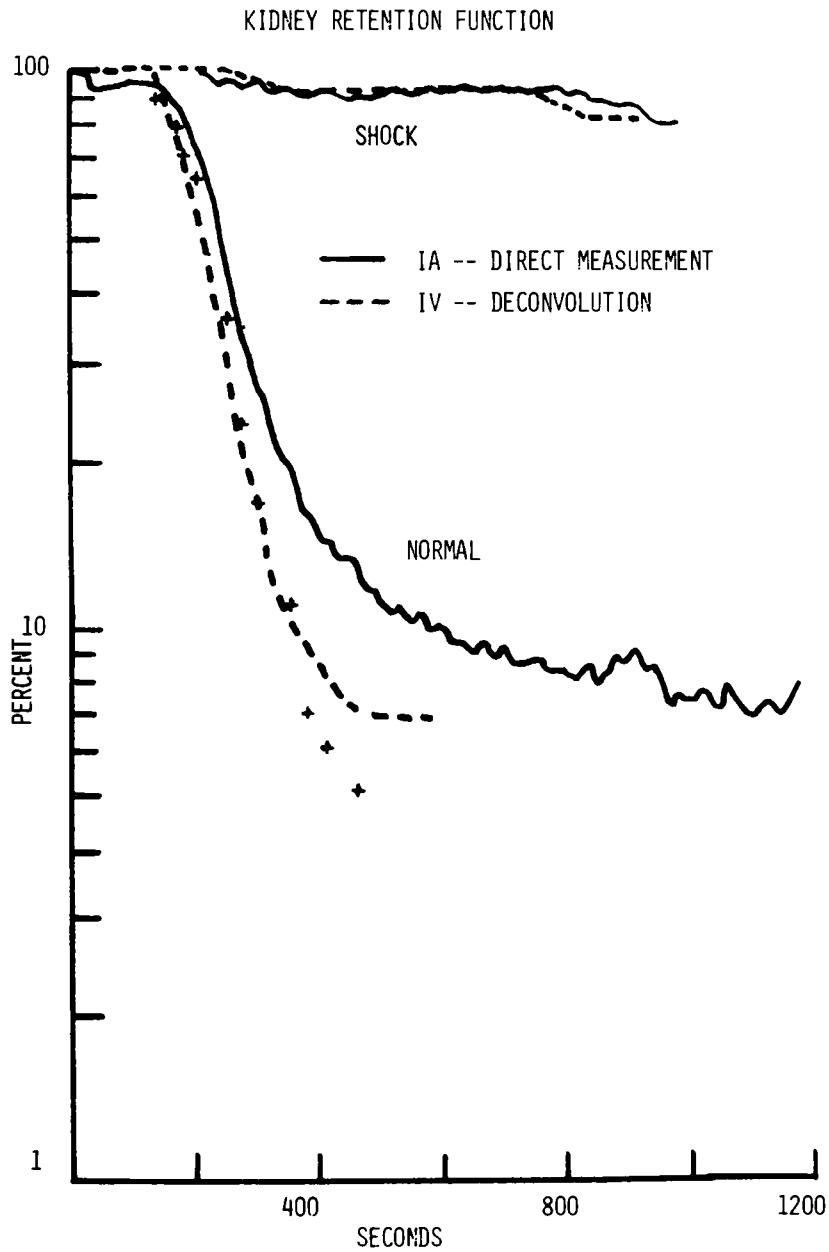


Figure 11. Direct measurements of renal retention function by means of intraarterial injection of I-131-OIH compared with deconvolved retention function from intravenous injection. T's correspond to deconvolved retention function corrected for extraction efficiency.

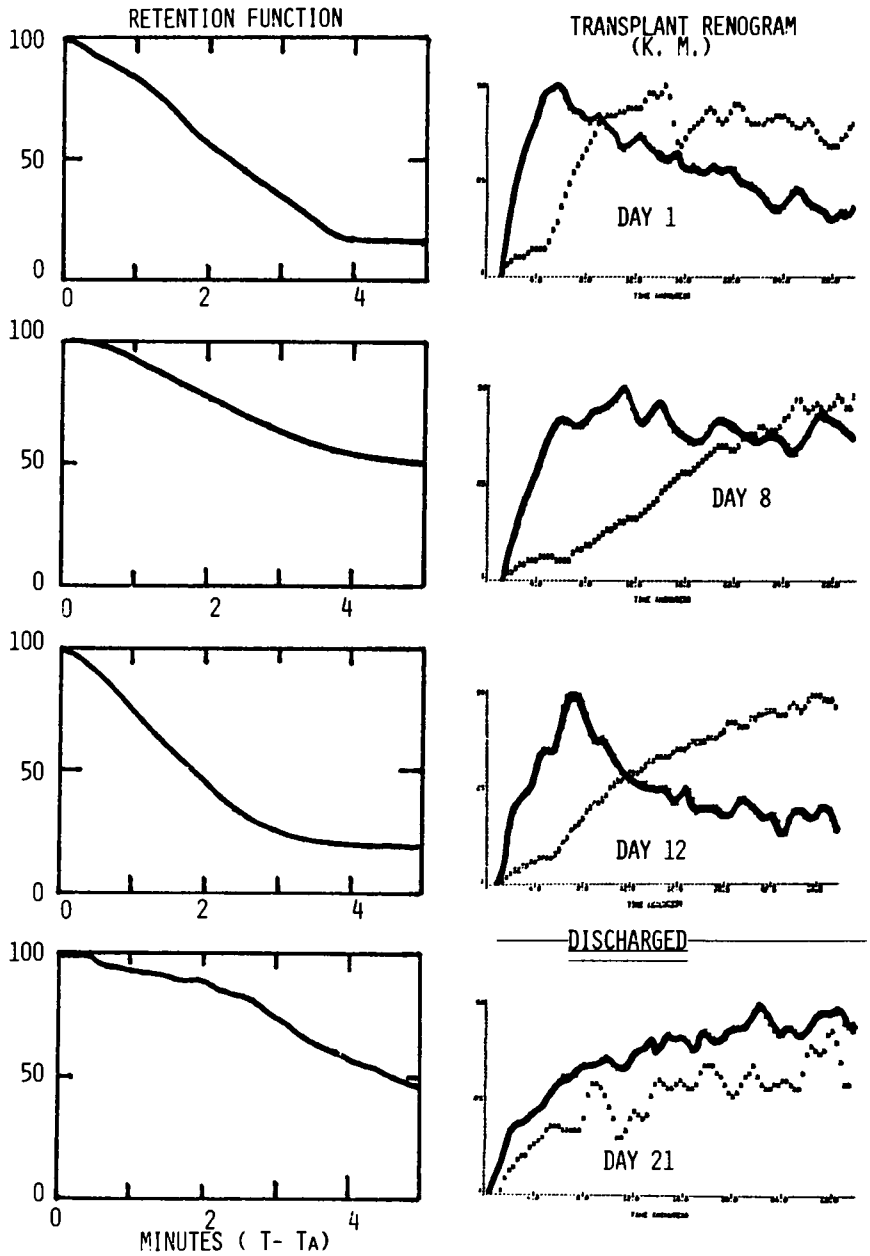


Figure 12. Sequential studies on a patient with a transplanted kidney. (Left) Deconvolved retention function days 1 through 21 after transplantation, (right) corresponding renogram.

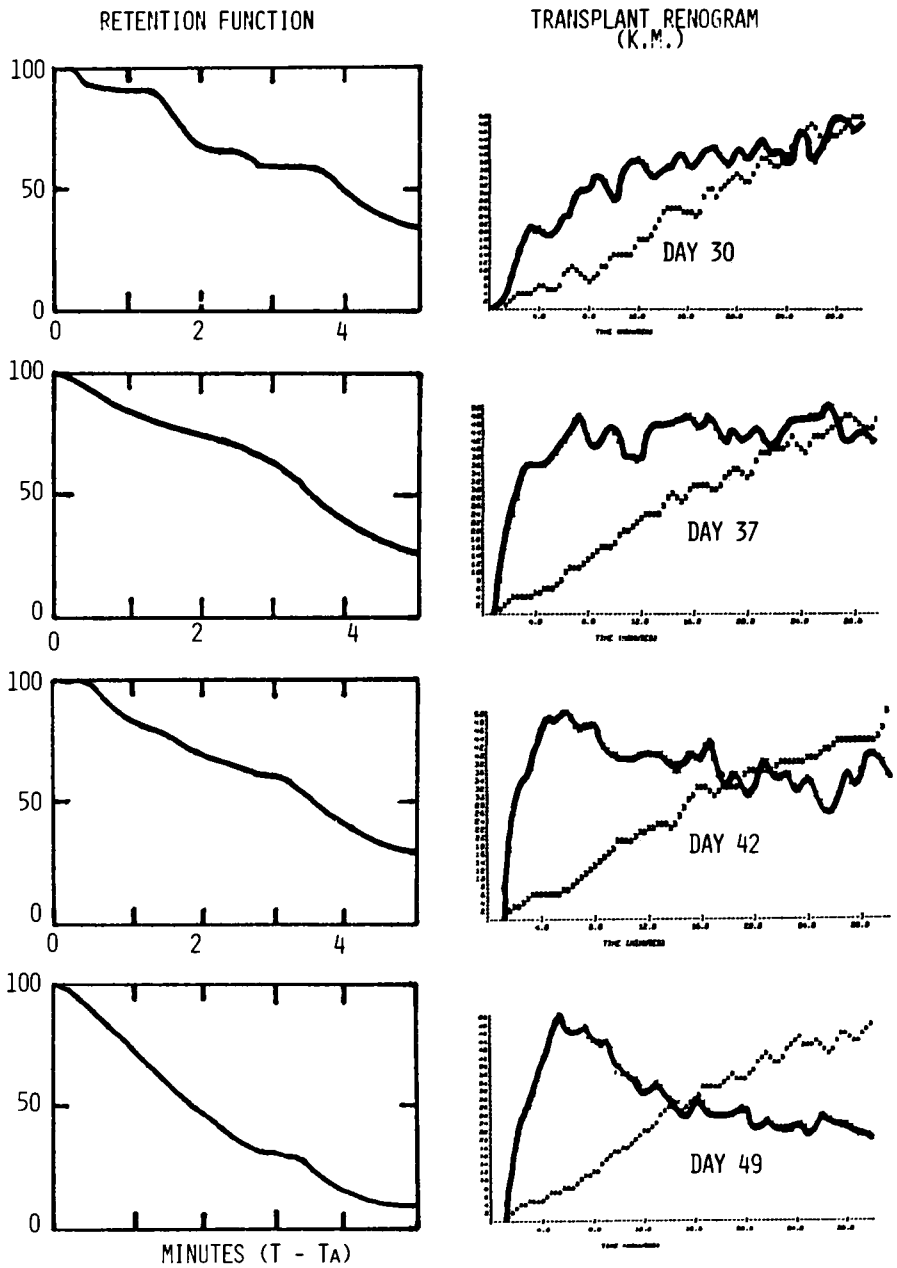


Figure 13. Continuation of Figure 12, days 30 to 49.

10). At this time an intensive course of both steroid and radiation therapy was begun. He began to respond to therapy, regaining good function and by day 49 was released from the hospital (BUN = 40, creatine = 1.9).

In general, as is the case in this example, the retention function curve is easy to interpret and is quite sensitive to changes in renal function. The retention function takes out differential rates of uptake from the kidney curves and enables one to view the rate of release of the compound from the kidney after trapping. However, the clinical utility of this technique relative to that of the routine renogram remains to be tested.

Summary and Conclusions

We have investigated two deconvolution techniques for estimating renal transit-time distributions. These were the Fast Fourier Transform technique and a point-by-point discrete deconvolution. In most situations both techniques yielded good results. However, we routinely used the point-by-point technique because of its increased speed of calculation, its insensitivity to "end effects" in the data and its relative insensitivity to high frequency components in the data.

Both deconvolution techniques were found to be extremely sensitive to the level of random noise in the data. Unprocessed data at count levels normally obtained in an I-131-OIH renogram gave uninterpretable results. Some form of data processing to reduce the effects of random noise was found to be essential. We arrived at a low frequency pass filter which attenuated frequencies with periods shorter than one minute. This filter was found to introduce an uncertainty in the determination of minimum transit-times less than one minute. This was not found to be a limitation in any of our animal or patient studies.

The presence of a blood background component in the renogram was found to also introduce an uncertainty in the minimum transit-time.

The deconvolution technique was unable to detect the presence of non-linearity in the system. Systems with reflux were interpreted as having an accelerated transit-time.

Direct measurements in experimental animals displayed the precision and accuracy of the deconvolution technique to be good over a wide range of renal function. We have seen no evidence for a bimodal structure in any of our transit-time curves. This observation includes in particular the intraarterial injections in dogs where the counting statistics were exceptionally good. All of our observations lead us to the conclusion that the normal kidney behaves very similiar to a classical mixing compartment plus a delay.

The relative clinical utility of the deconvolution technique with respect to the traditional renogram is not clear at this time.

REFERENCES

1. Britton, K.E., and Brown, N.J.G.: "The Clinical Use of C.A.B.B.S. Renography," Br. J. Radiology, 41, page 570, 1968.

2. Kenny, R.W., Ackery, D.M., Fleming, J.S., Goddard, B.A., and Grant, R.W.: "Deconvolution Analysis of the Scintillation Renogram," Br. J. Radiology, 48, page 481, 1975.

3. Britton, K.E., and Brown, N.J.G.: Clinical Renography, Lloyd-Lake, London, 1971.

4. Meldolesi, U., Roncari, G., Fidenga, M.A., et al.: "First Attempts of Clinical Application of a Method for the Quantitative Interpretation of Renogram with Radiohippuran." J. Nucl. Biol. Med. 13: p. 94, 1969.

5. Lindmo, T., Skretting, A. and Nakken, K.F.: "An Examination of Different Mathematical Models of Renal Function as Measured by ^{131}I-Hippuran Renography." Medical Physics, Vol. 1., No. 4., p. 193, 1974.

DISCUSSION:

Britton commented that in the transit time distributions presented he did not see the bimodality often observed. Britton stated that he felt that the shape of the distribution that he had re

Britton commented that in the transit time distributions presented he did not see the bimodality that he had often observed. Britton stated that he felt that the shape of the distribution was related to differences in cortical and juxtamedullary function.

Price noted that he had not seen multiple components in the transit time distribution in dogs. He noted further that they considered the possibility that these features were being filtered out, but stated that he was convinced that this was not the case. He stated that it appears their data can be described with a simple one compartment model with a delay.

Britton, in answering the questions arising about the validity of a technique, stated that he felt it imperative to evaluate and understand the factors influencing the distribution of transit times in the normal kidney. Otherwise, it should be very difficult to estimate the effects of noise and physiologic variation on the results in abnormal patients. He noted that several pieces of clinical and experimental evidence which seem to indicate that the distributions of transit time spectra do have clinical correlates. Britton described an early set of studies where sodium and potassium intake in a population of hypertensive patients was conducted. They studied the proportion of Hippuran traveling through the kidney with transit times under the first mode of a bimodal distribution. The proportion under the first mode in the state of low salt intake was over twice that in the state of high salt intake. The salt intake therefore does alter the distribution of transit times and is just as important to consider as the control of water intake. In another study, they measured plasma renin levels in patients with cirrhosis on a low salt diet. The renal function in these patients was normal; they all demonstrated creatinine clearances above 90 ml/min. The results showed that when the first or cortical component of the transit time distribution was large, the renin levels were normal. There was a significant negative correlation between the area under the first peak and the level of renin activity. Thus, he concludes that they have found that it is very important to control the various physiologic factors which can affect the transit time distributions before statements can be made about the effects of pathological influences.

Brill suggested that to resolve the apparent discrepancy between the shapes of the time distributions, it would be useful to determine if details of the distributions are being observed by mathematical manipulations. He suggested a comparison of the results computed from the same group of patient studies would be very helpful.

SOME NEW SIRT-TYPE RECONSTRUCTION ALGORITHMS

A.V. Lakshminarayanan and A. Lent
Medical Image Processing Group
Department of Computer Science
State University of New York at Buffalo
Amherst, N.Y. 14226 USA

ABSTRACT

We propose some new reconstruction algorithms which are modifications and extensions of SIRT. The least-squares theory underlying these algorithms is developed, and their convergence is proved. Experimental tests of the speed and stability of these algorithms in the presence and the absence of noise are presented.

Introduction

In a recent paper[1] we analyzed the SIRT algorithm (Gilbert[2]) and identified it as an approximate version of Richardson's least-squares algorithm for solving systems of linear equations. In this paper we present some generalizations and modifications of the SIRT algorithm which were suggested by our earlier analysis.

In Sections 2 and 3 the necessary vector and matrix notation is introduced, and the SIRT algorithm is presented in these terms. Section 4 summarizes pertinent results from LL. (We use the abbreviation LL for Lakshminarayanan and Lent[1]). In Section 5 a generalized, more rapidly converging version of the SIRT algorithm is introduced. Still more basic modifications of the SIRT algorithm are proposed in Section 6. Finally, we experimentally compare the performance of representative SIRT algorithms on various test pictures (Section 7).

Reconstruction Problem in Vector Notation

The algebraic methods for reconstruction make a number of physical assumptions and mathematical approximations. The ones made by Gilbert[2] are:

a. The projection data are collected from parallel rays of finite width w. The data are subject to experimental error.

b. An approximation to the true image is sought from the class of discrete images. A discrete image is a piece-wise constant function which may take on a different value over each square of an $n \times n$ grid of identical squares (pixels) of side h. The grid covers a region which contains the true image. When h is made small enough, there is a discrete image which resembles the true image

c. The projection data which would come from a discrete image are themselves approximated. The integrated density within a ray is approximated by the sum of all the density values of pixels whose centers fall within the ray (this sum is called a raysum) multiplied by h^2 (the area of a single pixel). This last approximation may be viewed as either an approximate integration formula for integrating along the true ray or else as an approximation of the true ray by one which coincides with pixel boundaries.

After making these approximations, it is convenient to introduce vector notation.

Let $N = n \times n$ be the number of pixels. Number these pixels in any (arbitrary but fixed) order. For any discrete image, let ρ_i be the density of the image in the i'th pixel. Define the image vector, ρ, as an N-component column vector whose i'th component equals ρ_i.

Let M stand for the number of projection data points (number of rays). Number these data in any (arbitrary but fixed) order. Let R_j stand for the j'th measurement. Define the measurement vector, R, as an M-component column vector whose j'th component equals R_j. For the j'th ray, we have, using (c), the approximate equality

$$h^2 \times (\text{sum over all pixels whose centers fall in the j'th ray}) \approx R_j. \tag{2.1}$$

We introduce some symbols in order to make this statement more compactly. Let

$$P_{ji} = \begin{cases} 1, & \text{if the center of the i'th pixel is contained in the j'th ray} \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

With this definition of P_{ji}, (2.1) may be rewritten as

$$h^2 \sum_{i=1}^{N} P_{ji}\rho_i \approx R_j. \tag{2.3}$$

For the final compaction, let P stand for the matrix whose element in the j'th row and i'th column is P_{ji}. The collection of M expressions of the form (2.3) may be written (making an obvious transposition of h^2) as

$$P\rho \approx R/h^2. \tag{2.4}$$

P is called the (unweighted) <u>projection</u> matrix. Equation (2.4) is called the <u>reconstruction equation</u>. Because of the errors introduced by (a), (b) and (c), an approximate, not an exact, equality is specified in (2.4)

Equation (2.4) is a restatement in vector notation of the reconstruction problem. Once we have solved (2.4) for ρ, we have reconstructed the unknown image from its projection data.

The SIRT Algorithm in Matrix Notation

<u>Standard Mathematical Notation</u> (See Halmos[3], or Ben-Israel & Greville[4])

(x,y) <u>Inner product</u> of the vectors x and y.

Let x and y be d-dimensional real vectors.

$$(x,y) \equiv \sum_{i=1}^{d} x_i y_i$$

$|x|$ Norm of the vector x.

Let x be a d-dimensional real vector.

$$|x| \equiv (x,x)^{\frac{1}{2}}.$$

$\underset{\sim}{R}(A)$ <u>Range</u> (or image space) of the matrix A.

$\underset{\sim}{R}(A) \equiv$ the set of all vectors y which can be written as $y = Ax$ for some vector x.

A^T <u>Transpose</u> of the matrix A.

Given a matrix A, $(A^T)_{ij} \equiv A_{ji}$.

$\lambda_{max}^{(A)}$ Largest eigenvalue of the matrix A.
We assume that A is a symmetric matrix

The SIRT Algorithm

SIRT is an iterative process and so assigns a new density value to each pixel at each step of the iteration.

Let ρ_i^q be the estimated value of the density of the i'th pixel after q iterative steps. Let S_i be the set of indices of all rays in which pixel i is included. Let N_j be the number of pixels intersected by the j'th ray. Let L_j be the length of intersection of the center line of the j'th ray and the reconstruction region. Let N_θ be the number of angles at which projection data are available. Let w be the width of a ray and let h be the length of a side of a pixel. (See Figure 1.) Following Gilbert[2], we assume that all rays have the same width. Herman & Rowland[5] have shown how to relax this assumption. Let

$$\rho_a = \frac{1}{N \cdot N_\theta \cdot h^2} \sum_{j=1}^{M} R_j .$$

ρ_a is an estimate of the average pixel density which is based on the experimental projection data.

The following remark is critical to our understanding of SIRT.
Remark: (LL, Equation (10)) For an arbitrary M-dimensional vector, v,

$$(P^T v)_i = \sum_{j \varepsilon S_i} v_j$$

It proves convenient to define two diagonal matrices, D_N and D_L, by

$$(D_N)_{ii} \equiv \sum_{j \varepsilon S_i} N_j, \quad i = 1,2,\ldots,N,$$

$$(D_L)_{ii} \equiv \sum_{j \varepsilon S_i} L_j, \quad i = 1,2,\ldots,N,$$

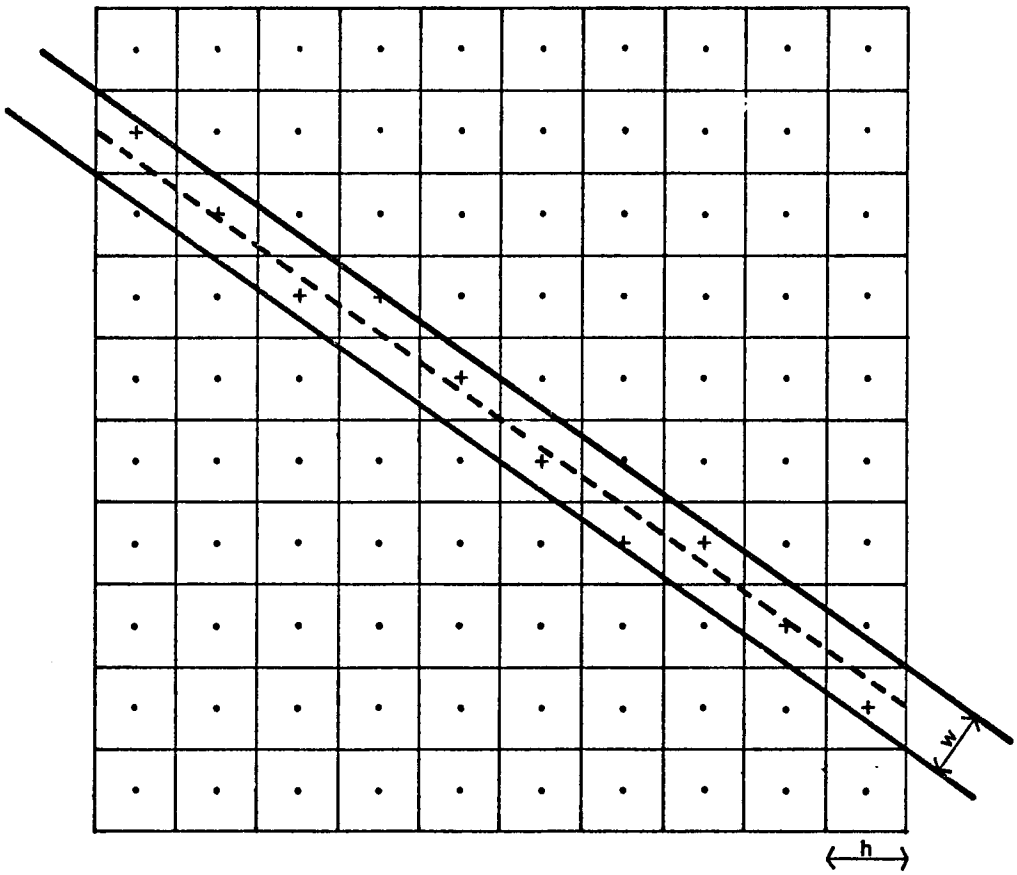
as well as a vector u:

$$u_i = 1, \qquad i = 1,2,\ldots,N.$$

In terms of these matrices, the SIRT algorithm takes the form

RECONSTRUCTION REGION

WITH TYPICAL RAY



• PIXEL CENTER

+ PIXEL CENTER WHICH IS CONTAINED IN RAY

--- CENTER LINE OF RAY

THE LENGTH OF THE CENTRAL LINE IN THE RECONSTRUCTION REGION IS "L"

Figure 1. The reconstruction region, showing pixels, of side h (small squares) and a typical ray (strip bounded by thick lines) of width w. Pixel centers are marked · and pixel centers which lie in the ray are marked +. The length of the central line of ray (---) in the reconstruction region is L.

$$\rho^{q+1} = \rho^q + (wD_L)^{-1} P^T R - D_N^{-1} P^T P \rho^q, \tag{3.1}$$

$$\rho^0 = \rho_a u.$$

This version of the SIRT algorithm (called unconstrained) generates reconstructions which are linear functions of the projection data. Other versions of SIRT (constrained, normalized) modify the algorithm presented so as to produce reconstructions which are everywhere nonnegative, and/or which have an average density equal to ρ_a.

Previous Results

In order to make the exposition of this paper reasonably self-contained, some pertinent results from LL are collected in this section.

a. Least Squares

Given a set of inconsistent equations

$$Ax \approx b, \tag{4.1}$$

the method of least squares looks for a vector x which makes $|Ax-b|$ as small as possible. If there is more than one such minimizing vector some auxiliary restrictions may be imposed.

b. Richardson's Algorithm

One of the standard methods for solving (4.1) is the Richardson iterative algorithm[6], which is stated as the following theorem:

Theorem: Let x^0 be any vector in $\underset{\sim}{R}(A^T)$. Let $\sigma > \frac{1}{2} \lambda_{max} (A^T A)$. Then the sequence $\{x^q\}$ generated by the iteration

$$x^{q+1} = x^q + \frac{1}{\sigma} (A^T b - A^T A x^q)$$

converges to a least-squares solution of

$$Ax \approx b.$$

If there is more than one least-squares solution, the sequence converges to the one of minimum norm.

Corollary 1: Let D be a positive-definite matrix. Let x^0 be any vector in $\underset{\sim}{R}(D^{-1}A^T)$. Let $\sigma > \frac{1}{2} \lambda_{max} (D^{-1}A^T A)$. Then the sequence $\{x^q\}$ generated by the iteration

$$x^{q+1} = x^q + \frac{1}{\sigma}D^{-1}(A^Tb-A^TAx^q)$$

converges to a least-squares solution of

$$Ax \approx b.$$

If there is more than one such solution, the sequence converges to the one which minimizes

$$(Dx,x)$$

c. SIRT

Lemma 1: Let u be the vector with $u_i = 1$, $i=1,\ldots,N$. u is a right eigenvector of $D_N^{-1}P^TP$ with corresponding eigenvalue 1.0.

Lemma 2: The largest eigenvalue of the matrix $D_N^{-1}P^TP$ is 1.0.

Theorem 2: A necessary and sufficient condition for the convergence of SIRT is the solvability of the equation

$$D_N^{-1}P^TP\rho = (wD_L)^{-1}P^TR. \tag{4.2}$$

If there is more than one such solution, the sequence $\{\rho^q\}$ generated by the SIRT iteration (3.1) converges to the solution which minimizes

$$(D_N\rho,\rho)$$

Generalizations of SIRT

In LL we have shown how, by a particular choice of relaxation parameter and starting vector, Richardson's algorithm may be identified with SIRT. It seems natural, then, to generalize SIRT so that it has the same variable parameters and the same potential for improved convergence that Richardson's algorithm does.

We define a generalized SIRT algorithm as follows:

$$\rho^{q+1} = \rho^q + \frac{1}{\sigma}((wD_L)^{-1}P^TR-D_N^{-1}P^TP\rho^q), \tag{5.1}$$

$$\rho^0 = \text{any advantageous starting picture.}$$

Like SIRT, this algorithm will converge only if Equation (4.2) is solvable. If this condition holds, the proof of Theorem 2 may be

extended to show that generalized SIRT converges if $\sigma > 1/2$ and $\rho^0 \in \underset{\sim}{R}(D_N^{-1}P^T)$.

For reasons given below, we suggest the use of the pair $\sigma=1/2$ and $\rho^0 = (wD_L)^{-1}P^TR$. This combination we call GSIRT.

Reasons for the choice $\sigma = 1/2$:

From Equation (5.1) we see that decreasing σ causes an increase in the size of the correction. This suggests the possibility of improving the rate of convergence by controlling σ; the question is how small σ can be made before the sequence of iterates develops increasing oscillations because of over-correction.

For the Richardson algorithm of Section 4b, the best value of σ to choose is

$$\sigma = 1/2(\lambda_{max}(D^{-1}A^TA) + \lambda^+_{min}(D^{-1}A^TA)),$$

where $\lambda^+_{min}(D^{-1}A^TA)$ stands for the smallest positive eigenvalue of $D^{-1}A^TA$. For the reconstruction problem, where D is replaced by D_N and A is replaced by P, we know from Lemma 2 that $\lambda_{max}(D_N^{-1}P^TP) = 1$. From a number of numerical experiments, the evidence is that $\lambda^+_{min}(D_N^{-1}P^TP)$ is several orders of magnitude smaller than $\lambda_{max}(D_N^{-1}P^TP)$. For these reasons, because of the close relationship between SIRT and Richardson's algorithm, $\sigma = 1/2$ seems like a good value to use in generalized SIRT.

Reasons for the choice $\rho^0 = (wD_L)^{-1}P^TR$:

The value $\sigma = 1/2$ is right on the border between convergence and oscillating divergence. Unless the starting point for the iteration is chosen carefully, the iteration will show an even-odd oscillation, and never converge. In this section we show mathematically how the oscillation comes about, and we suggest a way of removing it. We follow this with an experimental demonstration of the existence of the oscillation and the effectiveness of the proposed remedy.

Let v stand for the left eigenvector of the matrix $D_N^{-1}P^TP$ corresponding to the eigenvalue $\lambda = 1$. Lemma 1 showed that a right eigenvector of $D_N^{-1}P^TP$ corresponding to $\lambda = 1$ is the vector

$$v = D_N u.$$

As a temporary abbreviation, let t stand for $(wD_L)^{-1}P^T R$. Let us take the inner product of both sides of Equation (5.1) with the vector v. This gives

$$(\rho^{q+1},v) = (1 - \frac{1}{\sigma})(\rho^q,v) + \frac{1}{\sigma}(t,v).$$

Specializing to the value of interest, $\sigma = 1/2$, we get

$$(\rho^{q+1},v) = -(\rho^q,v) + 2(t,v).$$

From this equation it is easy to show that

$$(\rho^{2n},v) = (\rho^0,v) \qquad n = 1,2,\ldots$$

and

$$(\rho^{2n+1},v) = (\rho^1,v) = -(\rho^0,v) + 2(t,v), \; n = 1,2,\ldots$$

We conclude that the odd and the even reconstructions will behave differently unless ρ^0 is chosen so that

$$(\rho^0,v) = (t,v).$$

The easiest way to accomplish this is to choose

$$\rho^0 = t = (wD_L)^{-1}P^T R.$$

This choice is essentially a weighted back projection (Vainshtein[7]), and so has a certain intuitive appeal. Instead of an initial guess of a uniform picture, we begin with a reasonable first approximation and proceed to improve it by iteration.

In Figure 2 we plot the average reconstructed density as a function of the iteration number. (The reconstructions are of the Gilbert's CIRCLES of Section 7). Note the even-odd oscillation when the average starting point is used and the absence of such oscillations when the back-projected starting point is used.

Both the extensions of SIRT embodied in GSIRT, namely, the relaxation factor $(1/\sigma)$ and the back-projection starting point are important for a faster convergence. In support of this claim we present in Figure 3 graphs of residuals as a function of iteration number for the two σ's and the two starting points of SIRT and GSIRT.

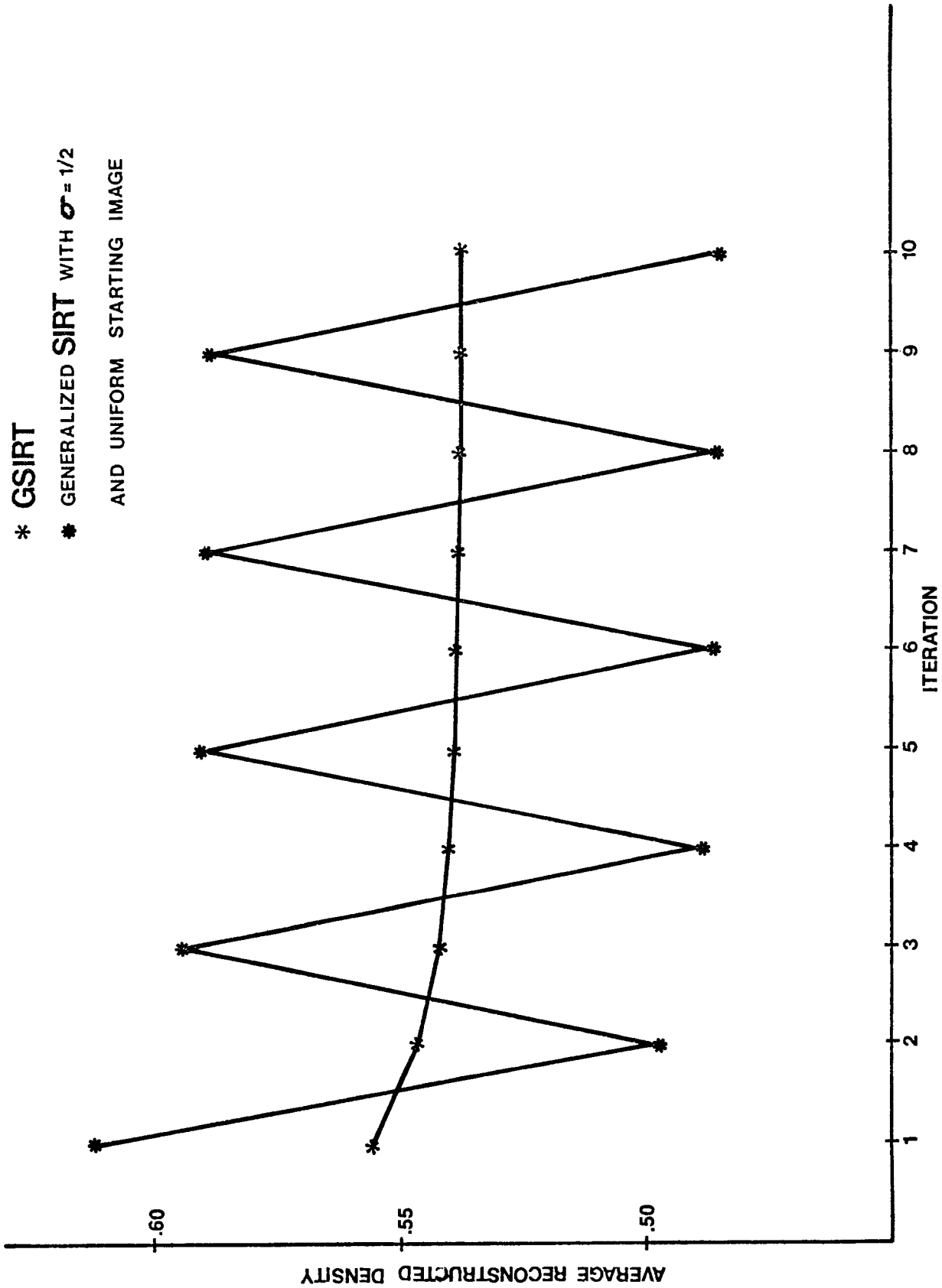


Figure 2. Plot showing the effect of the starting image on the reconstructed density as a function of iteration number. The generalized SIRT algorithm is used.

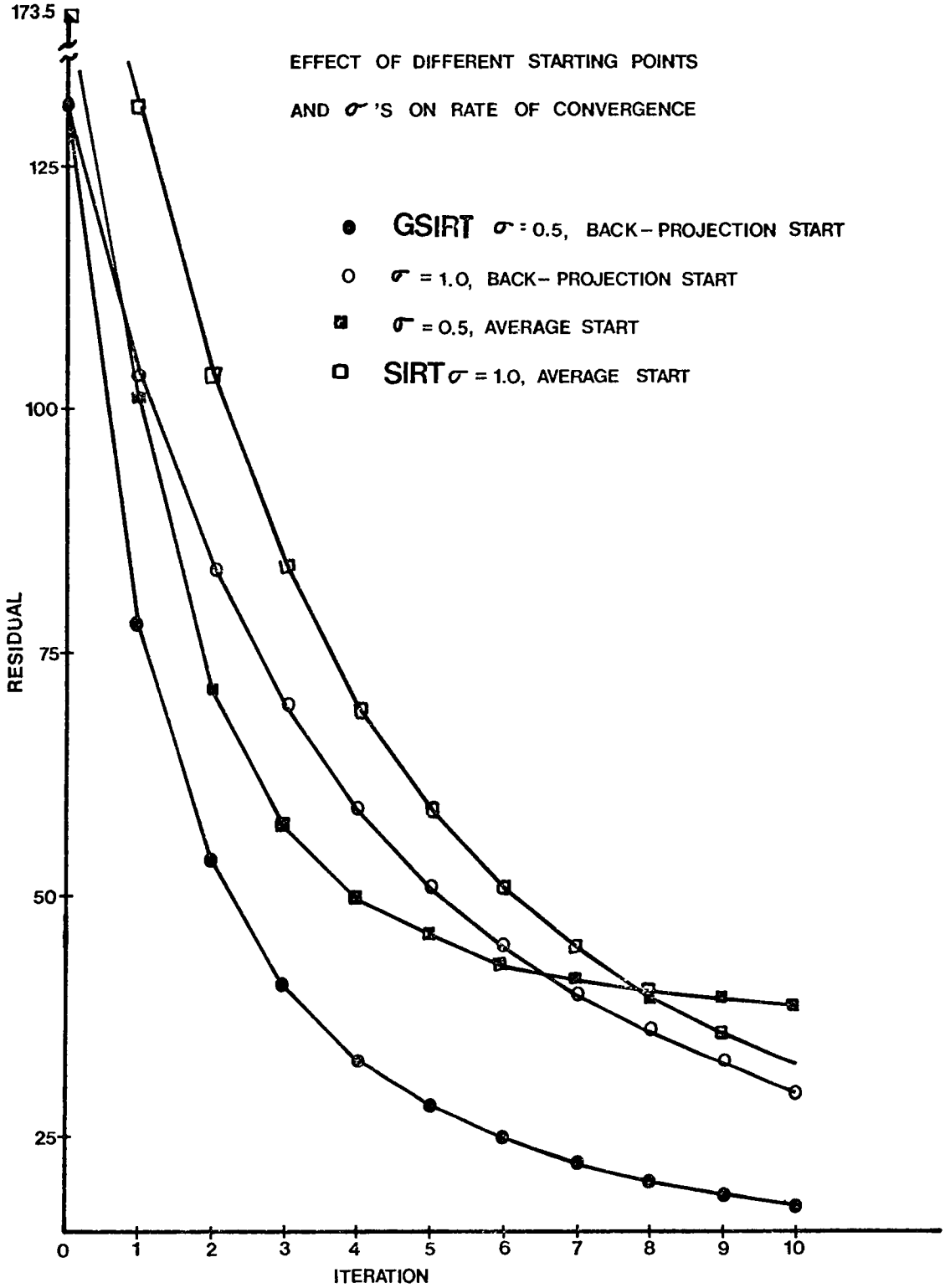


Figure 3. A plot of the sequences of residuals showing the importance of choosing both the starting picture and the relaxation factor σ properly.

□ SIRT: ○ σ = 1·0, back-projection start;
■ σ = 0·5, average start; ● GSIRT.

Note that GSIRT with both σ = 1/2 and back-projection starting picture has the smallest and most rapidly converging residuals.

We see that changing the starting point alone or the relaxation factor alone is ineffective, or even detrimental, but the combination produces a striking improvement.

Further tests of GSIRT are presented in Section 7.

Modifications of SIRT

SIRT makes a strong attempt to cope with two of the errors mentioned in Section 2. As has been shown in LL, Section 6, SIRT treats the data error (a) by means of least squares (approximately). Gilbert also attempts to reduce the effect (c), the error caused by approximate integration along rays. It is in order to reduce this error that the matrix wD_L appears in the SIRT algorithm. When R is the projection data from a uniform picture, i.e., $R_j = wL_j$, $i = 1,\ldots,M$, it can be seen from Equation (5.1) that SIRT will reproduce that picture almost exactly.

In this section, we propose a family of modified SIRT algorithms which have the above theoretical advantages, but which are true least-squares algorithms. In addition, the members of this family are all convergent algorithms. We call this family Least-Squares SIRT (LSIRT). The form of these algorithms is suggested by the close resemblance between SIRT and Richardson's algorithm.

LSIRT is based on the set of equations

$$\frac{L_j}{N_j} w_j (P\rho)_j \approx R_j, \quad j=1,\ldots,M, \tag{6.1}$$

which we attempt to solve in a weighted least-squares sense. (w_j stands for the width of the j'th ray. It is introduced to permit the use of variable ray widths.) These equations are implicit in Crowther & Klug[8] and in Gilbert[2]. Note that the left-hand side of Equation (6.1) is a quadrature formula for approximating the raysums which is more accurate (it does at least integrate constant functions correctly) than that of Equation (2.1).

The use of a least-squares fit to data is justified by the argument of Gauss (see Whittaker & Robinson[9], Section 115) assuming independent and identically distributed errors in the projection

data. If noise in the data varies (due to counting statistics, for example) a weighted least-squares fit (see Goitein[10]) is appropriate. Alternatively, a weighted least-squares fit may be used for reasons of numerical expediency. Usually, the solution is only slightly affected by changes in the weights.

Let z_j, $j = 1,\ldots,M$ be the given positive weights. A weighted least-squares solution minimizes

$$\sum_{j=1}^{M} z_j \left(R_j - \frac{L_j w_j}{N_j}(P\rho)_j\right)^2.$$

Let V stand for the M×M diagonal matrix with elements

$$V_{jj} = \frac{L_j w_j}{N_j}, \quad j=1,\ldots,M,$$

and let Z stand for the M×M diagonal matrix with

$$Z_{jj} = z_j, \quad j=1,\ldots,M.$$

Then, it may be shown that the solution to the weighted least-squares problem satisfies the equation

$$P^T Z V^2 P\rho = P^T Z V R. \tag{6.2}$$

We now develop a SIRT-like method which finds a solution of (6.2).

The feature which distinguishes a SIRT iteration from a Richardson iteration for solving (6.2) is the presence of a diagonal matrix multiplying the usual Richardson correction. We consider, therefore, an algorithm of the form

$$\rho^{q+1} = \rho^q + \frac{1}{\sigma} D_V^{-1}(P^T Z V R - P^T Z V^2 P\rho^q), \tag{6.3}$$

where D_V is a diagonal matrix chosen to make

$$\lambda_{max}(D_V^{-1} P^T Z V^2 P) = 1.$$

By the same reasoning as that used in Lemma 1 and Lemma 2, this will be achieved if

$$D_V^{-1} P^T Z V^2 P u = u,$$

which happens if

$$(D_V)_{ii} = \sum_{j \in S_i} z_j L_j^2 w_j^2 N_j^{-1}. \tag{6.4}$$

Using Corollary 1 on the convergence of Richardson's algorithm, we may prove convergence of LSIRT in the same way that the convergence of SIRT was proved.

Theorem 3: If $\sigma = 1$ and $\rho^0 \equiv \rho_a u$, then the iteration (6.3) converges to the solution of Equation (6.2) which minimizes $(D_V\rho,\rho)$.

Proof: The proof is the same in form as that of Theorem 2. The matrix $Z^{\frac{1}{2}}VP$ replaces P, and D_V replaces D_N,

The idea which was used in Section 5 to generalize SIRT may be used to generalize LSIRT as well.

Remark: If $\sigma = 1/2$ and $\rho^0 = D_V^{-1}P^T ZVR$, then the iteration (6.3) converges to the solution of Equation (6.2) which minimizes $(D_V\rho,\rho)$.

Choice of the matrix Z:

The choice of the weights (Z_{jj}) to be used in the least-squares scheme influences both the performance of the algorithm in the presence of measurement errors and the complexity of the computations. Here we have chosen two different weighting schemes which are very attractive from the computational point of view. These special choices of the Z-matrix and the designation of the resulting algorithms in our generalized SIRT routines are given below.

LSIRT1:

$$Z_{jj} = V_{jj}^{-2}, \quad j = 1,\ldots,M \tag{6.5}$$

$$(D_V)_{ii} = (D_N)_{ii}, \quad i = 1,\ldots,N$$

With this choice, Equation (6.3) specializes to

$$\rho^{q+1} = \rho^q + \frac{1}{\sigma} D_N^{-1}(P^T V^{-1} R - P^T P \rho^q) \tag{6.6}$$

LSIRT2:

$$Z_{jj} = N_j^{-1} V_{jj}^{-2}, \quad j=1,\ldots,M \tag{6.7}$$

$$(D_V)_{ii} = N_\theta, \quad i = 1,\ldots,N.$$

With this choice, Equation (6.3) becomes

$$\rho_i^{q+1} = \rho_i^q + (\sigma N_\theta)^{-1} \sum_{j \varepsilon S_i} \left(\frac{R_j}{L_j w_j} - \frac{(P\rho^q)_j}{N_j} \right), \tag{6.8}$$

$$i = 1,\ldots,N.$$

The choices of D_V in Equations (6.5) and (6.7) satisfy Equation (6.4). It is clear that these schemes remove the need for the calculation of L_j, $j=1,\ldots,M$, during the back-projection step of Equation (6.3). In LSIRT2 the matrix D_V is reduced to a scalar, making it quite attractive computationally.

We compare the performance of GSIRT, LSIRT1 and LSIRT2, and SIRT in the next section.

Comparison of the Different SIRT Methods

We have run three different series of tests on the different SIRT methods. In the first series we have reconstructed images from noiseless projection data. In the second, we have reconstructured from projection data with simulated noise, and in the last we have obtained "reconstructions" of pure noise.

We have used computer-printer-produced pictures of reconstructed images in each case to get a visual estimate of the goodness of the methods. Some of these pictures are given in this section.

In addition to the visual picture, we have used quantitative measures of the 'goodness' of the reconstructions for comparison purposes. These are

(i) The <u>residual</u> at each iteration

$$r^q \equiv \left(\sum_{j=1}^{M} (R_j - (P\rho^q)_j)^2 \right)^{\frac{1}{2}}. \tag{7.1}$$

(ii) The <u>average</u> reconstructed density

$$\bar{\rho}^{\,q} \equiv \frac{1}{N} \sum_{i=1}^{N} (\rho_i^q)$$

(iii) The variance is denoted v^q;

$$v^q \equiv \frac{1}{N} \sum_{i=1}^{N} (\rho_i^q - \bar{\rho}^q)^2$$

(iv) The distance, δ^q, measures the deviation of the reconstruction from the digitized test picture, ρ^t;

$$\delta^q = \left| \frac{1}{Nv^q} \sum_{i=1}^{N} (\rho_i^t - \rho_i^q)^2 \right|^{\frac{1}{2}} \tag{7.2}$$

Test Pictures:

The different versions of SIRT have been tested on two mathematically created phantoms, (a) CIRCLES and (b) SKULL, shown in Figure 4.

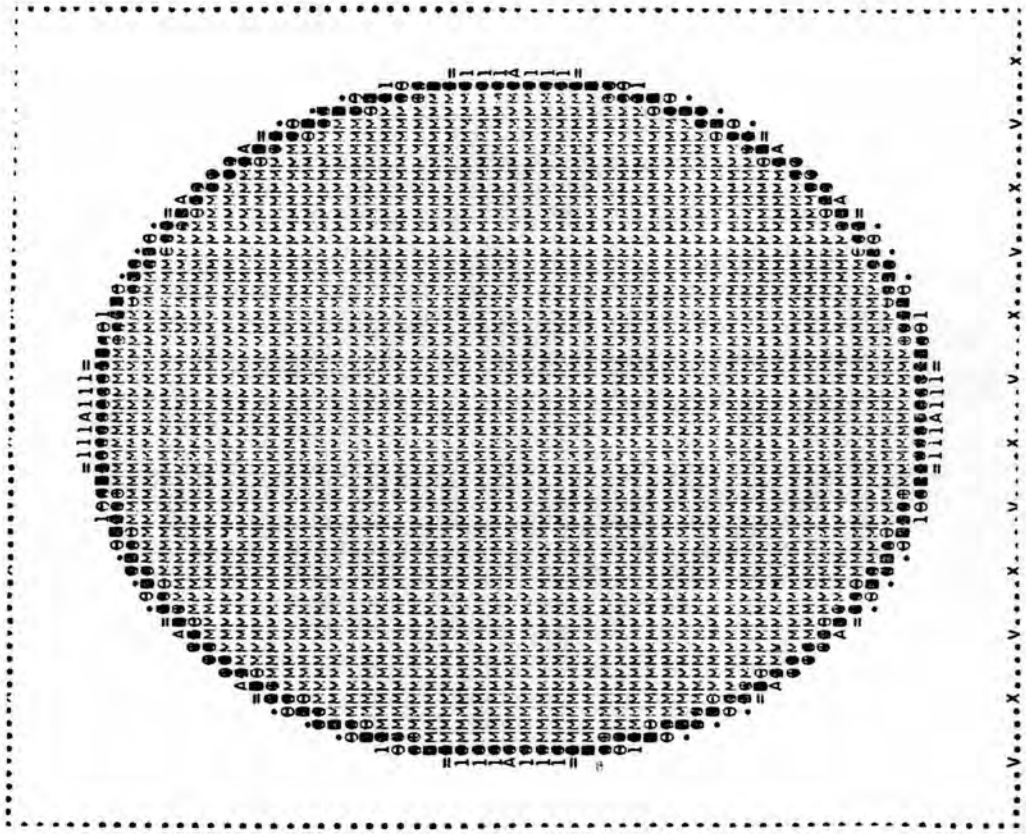
(a) CIRCLES: This is a 27×27 digitized version of the test patterm used by Gilbert[2], which is composed of five circles of density 1.5 in a circular region of density 0.5. Projection data have been generated (as described in Herman & Rowland[5]) at 25 angles equally spaced over 180° using parallel rays of width equal to the side of a picture element in the phantom. For each projection there are 41 rays.

(b) SKULL: This is a 63×63 digitized version of the pattern used by Herman & Rowland[5]. It is composed of an annulus 1 pixel thick with density 1.0 surrounding a circular region of radius 26 pixels, density 0.5. Projection data have been generated at 30 angles equally spaced over 180° using parallel rays of width equal to the side of a pixel. For each projection there are 93 rays.

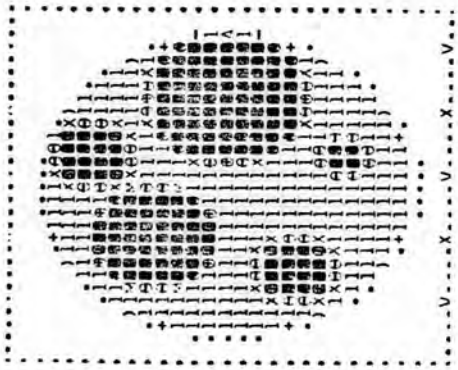
Noiseless Projection Data:

Figure 5 shows the reconstructions of SKULL produced by ten iterations of SIRT and GSIRT. The superiority of GSIRT is evident. GSIRT reconstructs the skull, while SIRT loses it completely (see also Herman & Rowland[5]).

Figure 6 shows plots of residuals produced when the algorithms GSIRT, LSIRT1, and LSIRT2 reconstruct the CIRCLES pattern. There is a close correspondence between the reconstructions produced by GSIRT and LSIRT1. LSIRT2 seems to converge faster than GSIRT (which is

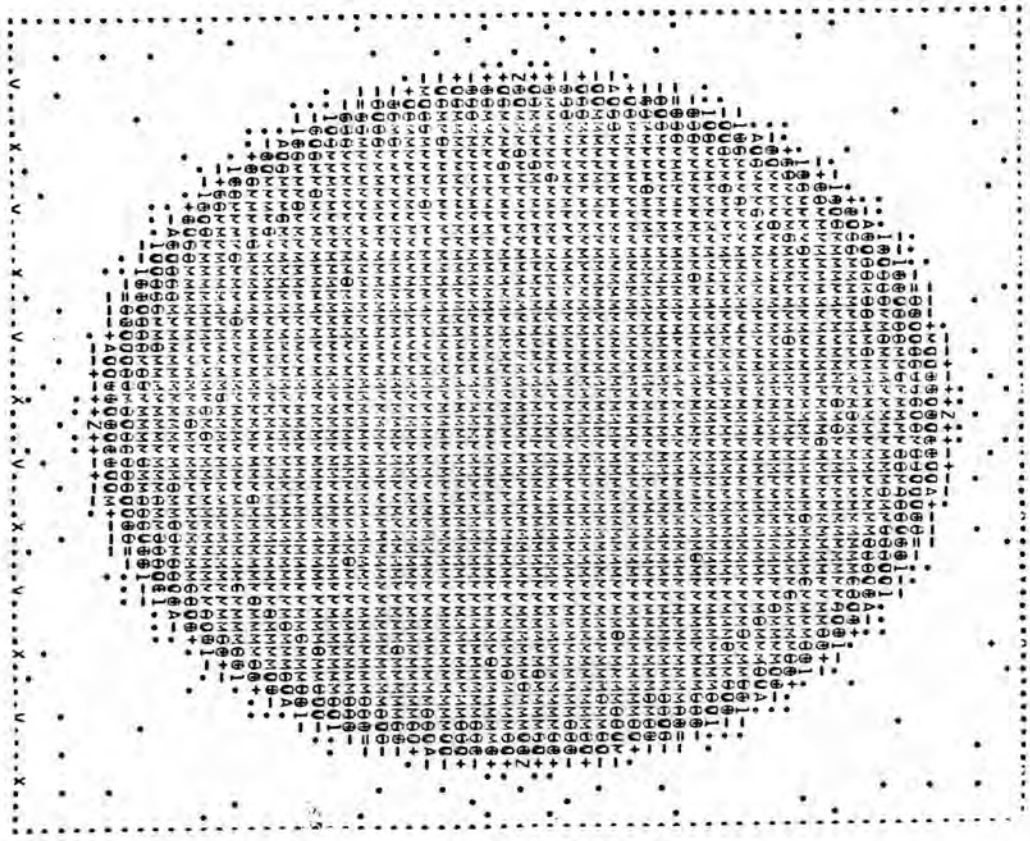


4b

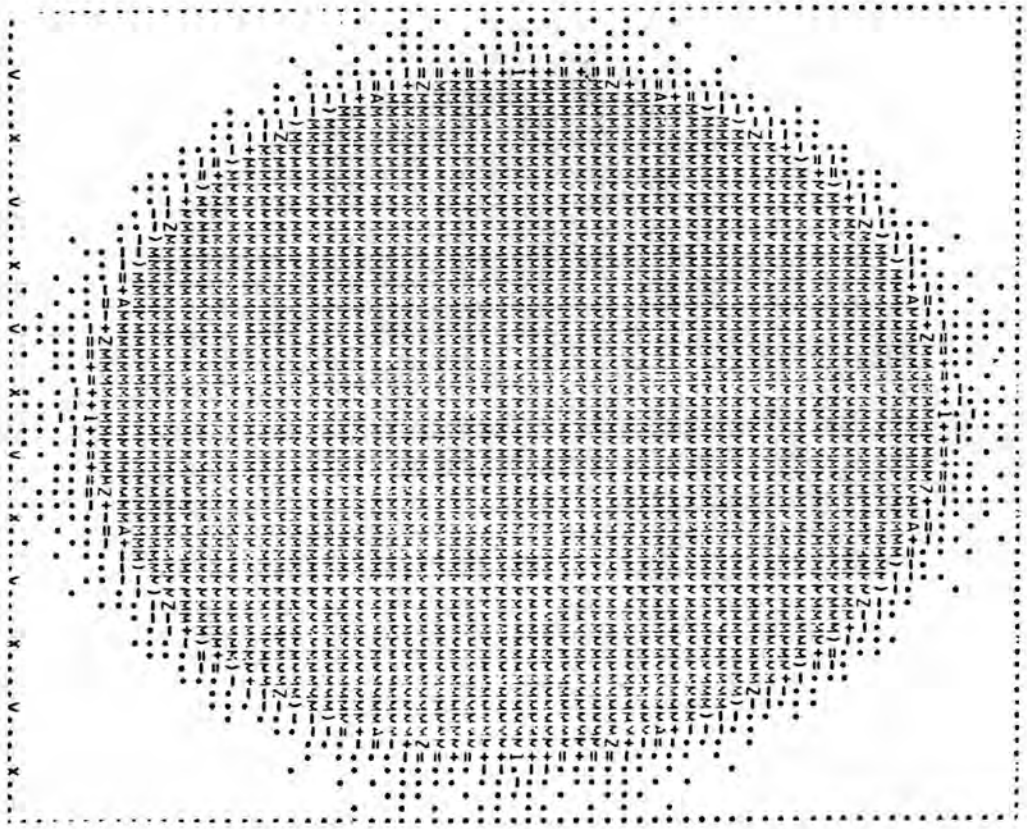


4a

Figure 4. A computer-produced display of the two phantoms used in our reconstructions.
(a) CIRCLES. A 27 × 27 version of the model object used by Gilbert.
(b) SKULL. A 63 × 63 version of the pattern used by Herman and Rowland.



5b



5a

Figure 5. A computer printer display of the reconstruction of the SKULL of Fig. 4(b) produced by ten iterations of (a) SIRT and (b) GSIRT. Note that SIRT loses the skull (outer dense ring of Fig. 4(b)) completely while GSIRT reconstructs it.

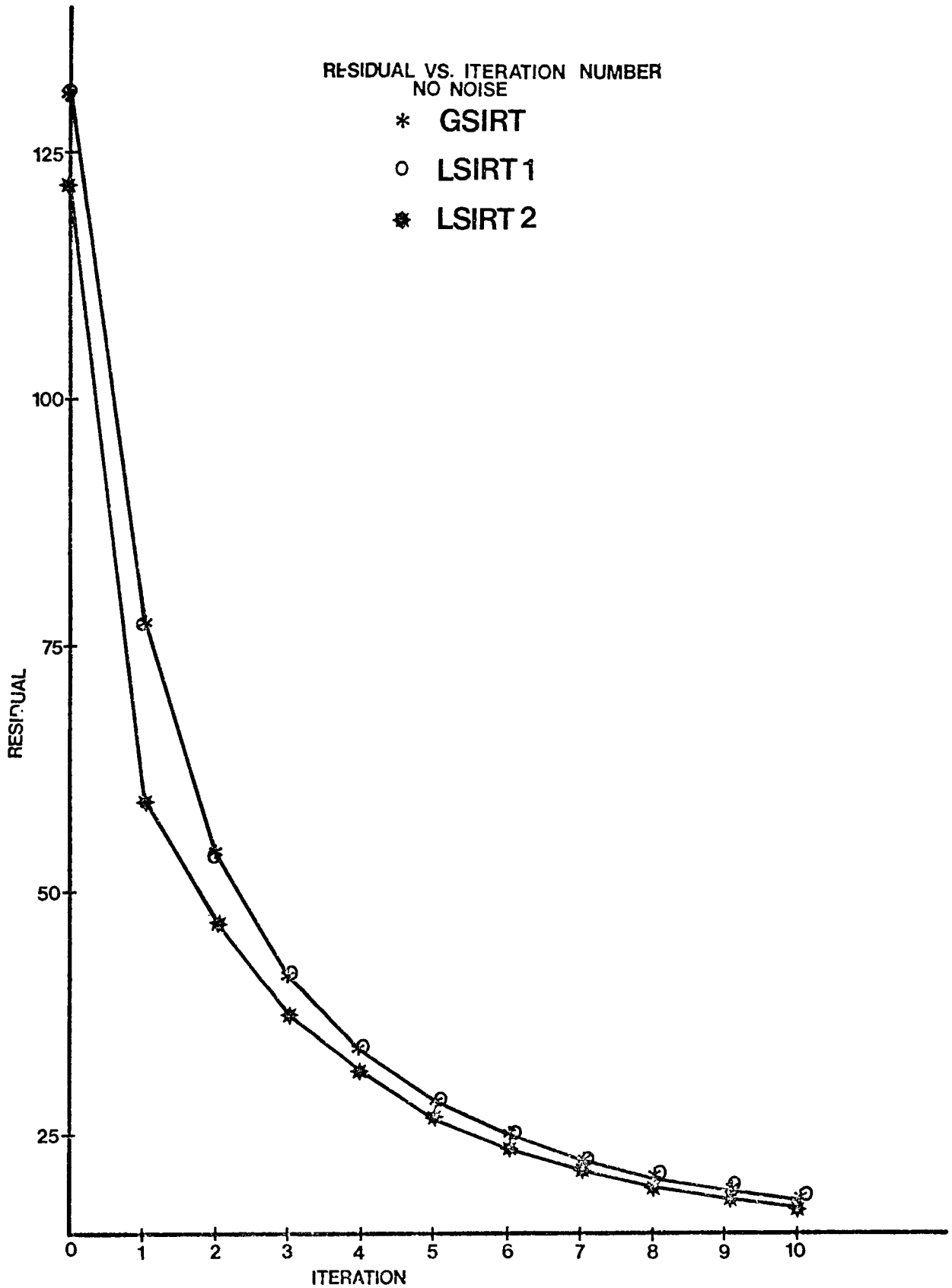


Figure 6. A plot of residuals r^q (of Equation 7.1) versus iteration number q for reconstructions of CIRCLES. Noiseless data.
* GSIRT, ○ LSIRT1, ● LSIRT2.
Note the closeness of GSIRT and LSIRT1.

itself faster than SIRT; see Figure 3), particularly at the beginning. Because there was no significant difference between GSIRT and LSIRT1 reconstructions only LSIRT2 was tested further. In Figure 7, pictures of CIRCLES produced by ten iterations of SIRT, GSIRT, and LSIRT2 are shown. The SIRT picture seems lacking in details and blurred compared to the other two.

On the basis of the SKULL and CIRCLES reconstructions it appears that, for noiseless data, our modifications of SIRT result in definite improvements.

Noisy Projection Data:

Noisy projection data were generated by multiplying the noiseless CIRCLES projection data by random samples from a Gaussian distribution with unit mean and standard deviation 0.1. (This experiment repeated Gilbert's original tests as closely as possible.) The performance of SIRT, GSIRT, and LSIRT2 reconstructing these data was examined. Some of the results are presented here.

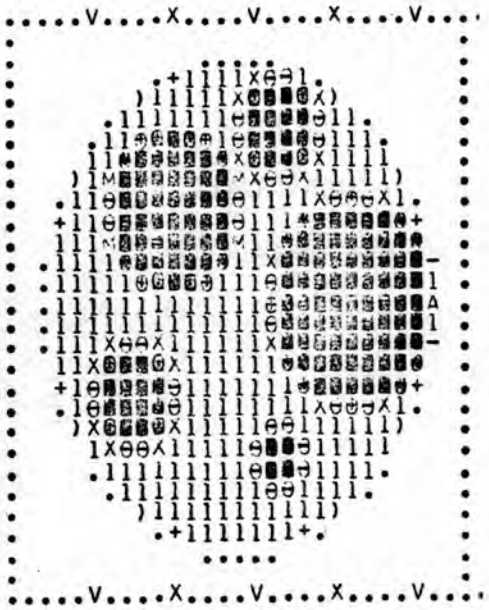
Figure 8 plots residuals versus iteration number; Figure 9 plots distance versus iteration number. Both the constrained and unconstrained versions of SIRT and GSIRT are presented in these figures.

Figure 8 shows that the residual decreases steadily, with GSIRT and LSIRT2 well ahead of SIRT. However, Figure 9 shows that, after a certain point, the quality of the GSIRT and LSIRT2 reconstructions is slightly degraded by further iteration. The reconstructed pictures also show this degradation in quality. The use of the unconstrained algorithms may degrade the picture more than the constrained versions.

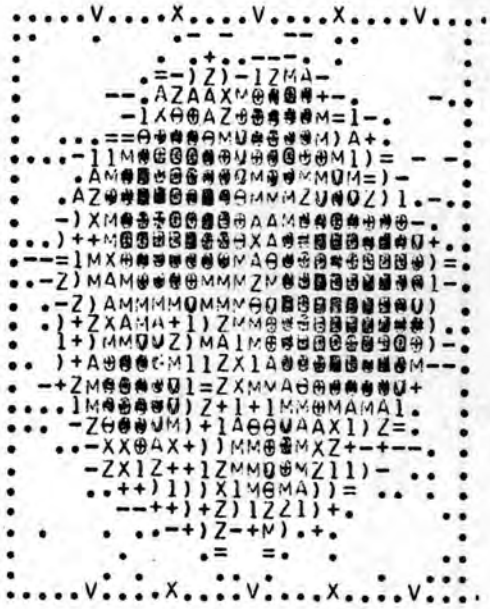
Thus, for the GSIRT and LSIRT2 methods in the presence of noise, the residual decreases steadily even as the reconstruction moves away from the test pattern. To explain this, the third series of experiments in which pure noise was used as projection data was run.

Pure Noise as Projection Data:

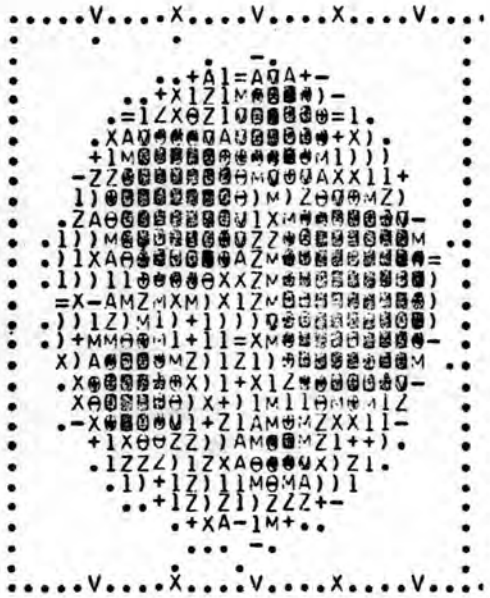
In order to separate the action of an algorithm on noise from



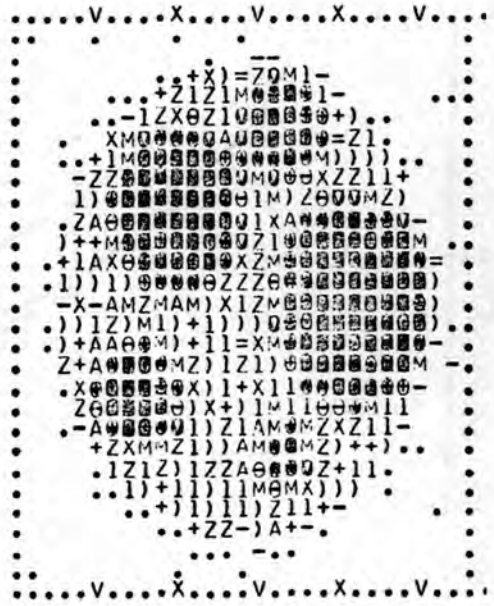
Phantom



SIRT



LSIRT2



GSIRT

CIRCLES Phantom and reconstructions at iteration 10 by SIRT methods

Figure 7. A computer printer display of the reconstruction of CIRCLES produced after 10 iterations of (a) SIRT, (b) GSIRT, and (c) LSIRT2.
Note the somewhat washed-out reconstruction obtained with SIRT.

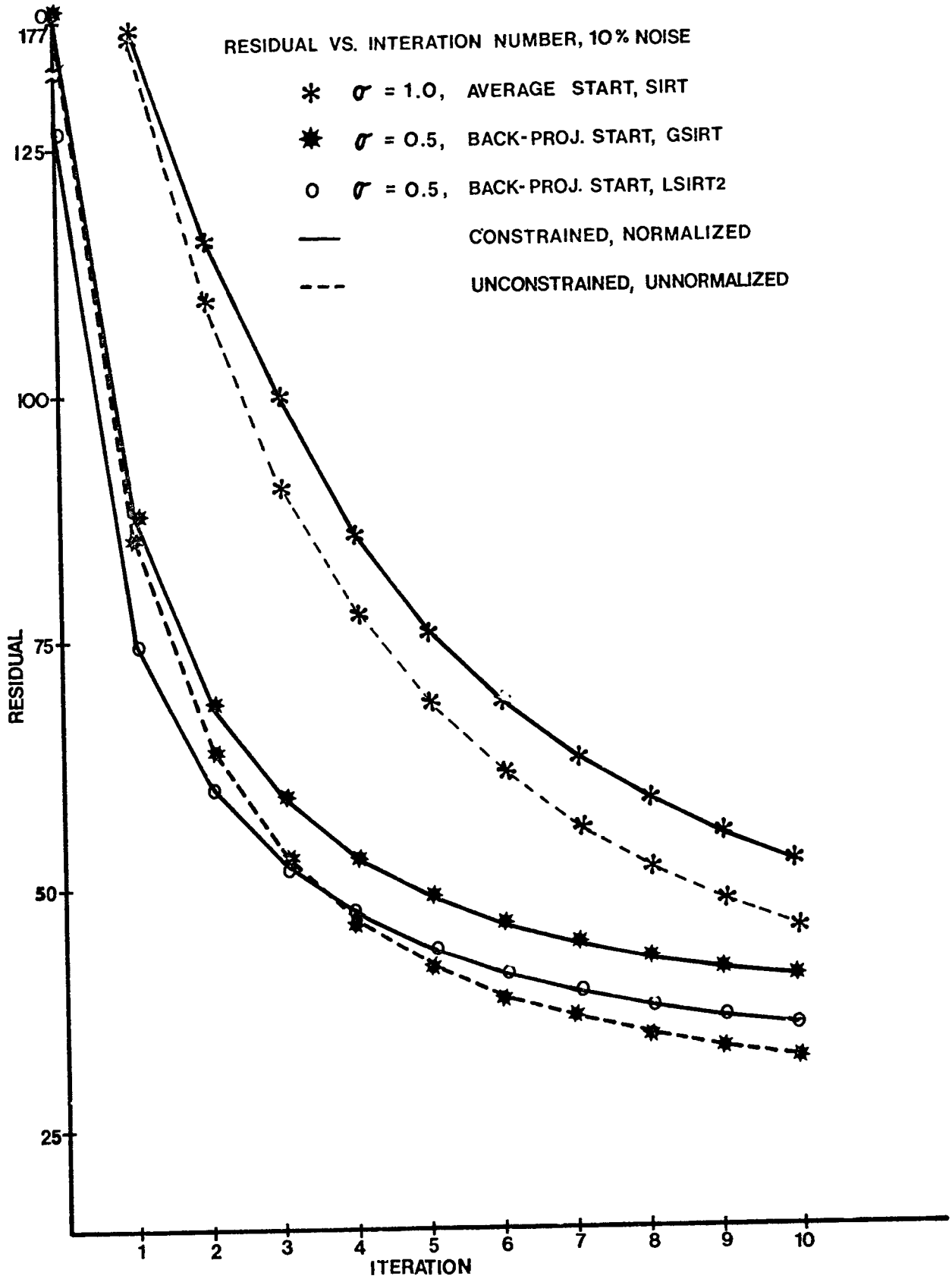


Figure 8. A plot of residuals r^q (of Equation (7.1)) versus iteration q for CIRCLES when 10% random multiplicative noise is introduced in the projection data.
* SIRT, ● GSIRT, ○ LSIRT2. Constrained values are shown connected by continuous lines and unconstrained values by broken lines.
Note that the residual decreases steadily, and GSIRT and LSIRT2 perform better than SIRT.

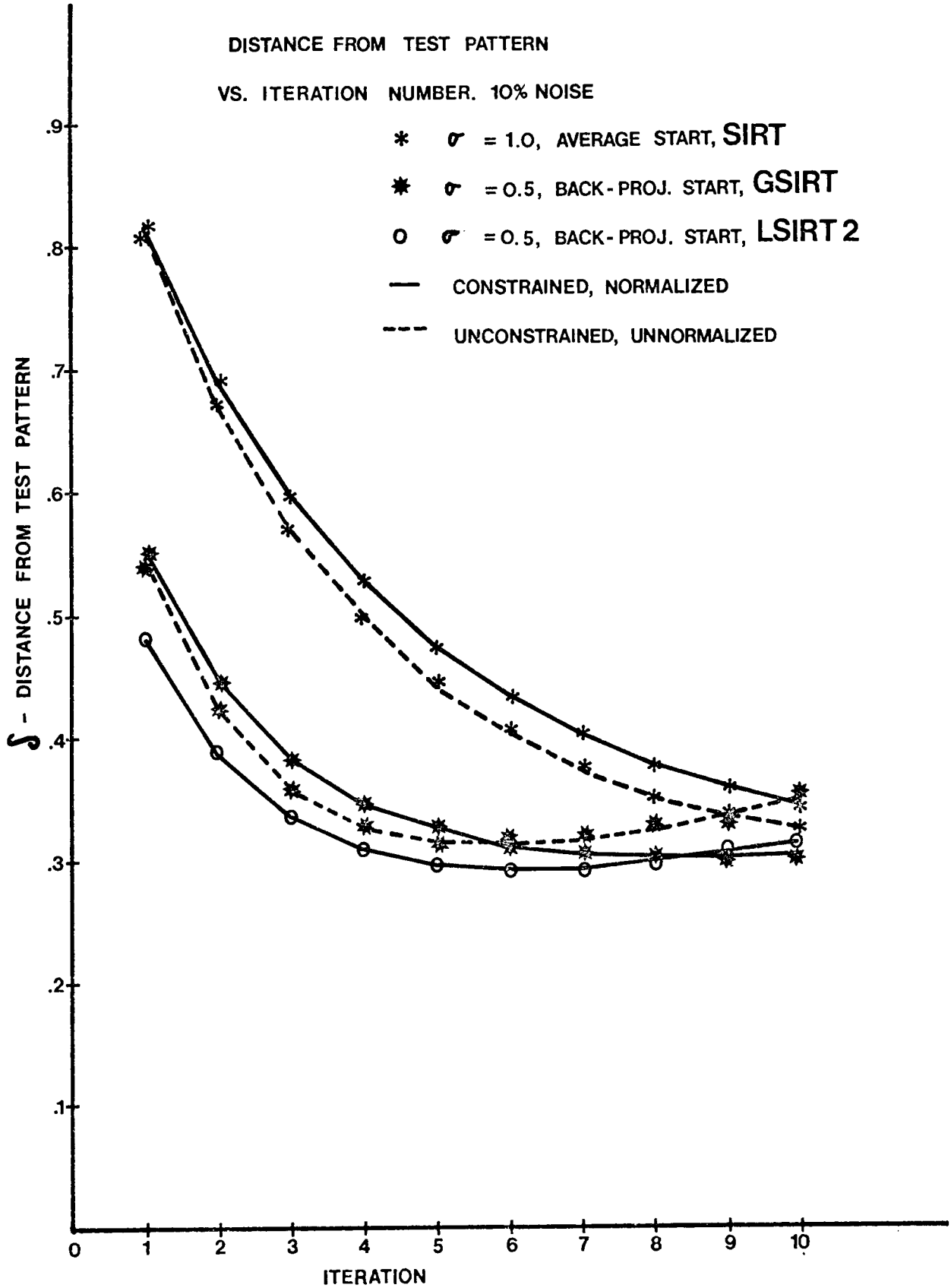


Figure 9. A plot of the distance δ^q (Equation 7.2) between the test picture (CIRCLES) and the q'th reconstruction. Symbols are the same as in Figure 8.
Note that after iteration 6 GSIRT and LSIRT2 tend to produce reconstructions which diverge from the test picture.

its action on data, it is simpler to assume the noise additive rather than multiplicative, as was done in the second series of experiments.

To simulate additive noise, samples are drawn from a Gaussian distribution with zero mean and standard deviation 1.0. These random numbers are given to an algorithm in place of projection data. Essentially, the algorithm then tries to reconstruct a zero (blank) picture for which the projection data supplied suffer from noise. Whatever image the algorithm produces is then pure error, due only to noise. By examining this image we learn about the sensitivity of the algorithm to noise. The use of this method was prompted by its successful use by Stuart Rowland[11] in his studies on convolution reconstruction methods. These ideas are now stated more precisely.

If constraining and normalization are not used, the SIRT algorithms generate reconstructions which are <u>linear</u> functions of the projection data. We can write

$$\rho^q = L^q R, \tag{7.3}$$

where L^q is the operator which converts the projection data into the q-th approximate reconstruction.

Now suppose there is random noise in the projection data. Let n_j be the random noise in the j'th measurement. Let R_j^e be the exact measurement which would have been made in the absence of noise. Then

$$R_j = R_j^e + n_j.$$

R_j^e and n_j can be thought of as the j'th component of the M-dimensional vectors R^e and n. Using Equation (7.3), we get

$$\rho^q = L^q R^e + L^q n.$$

Let ${}_e\rho^q \equiv L^q R^e$, the reconstruction from exact data, and let ${}_n\rho^q \equiv L^q n$, the reconstruction from pure noise.

We focus our attention on the summation appearing in the expression for δ^q in Equation (7.2):

$$\frac{1}{N}\sum_{i=1}^{N}(\rho_i^q - \rho_i^t)^2 = \frac{1}{N}\sum_{i=1}^{N}({}_e\rho_i^q + {}_n\rho_i^q - \rho_i^t)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}({}_e\rho_i^q - \rho_i^t)^2 + \frac{1}{N}\sum_{i=1}^{N}({}_n\rho_i^q)^2$$

$$+ \frac{2}{N}\sum_{i=1}^{N} {}_n\rho_i^q\,({}_e\rho_i^q - \rho_i^t).$$

When we take the expected values over the random variable n, the last term on the right drops out.

$$E[\frac{1}{N}\sum_{i=1}^{N}(\rho_i^q - \rho_i^t)^2] = \frac{1}{N}\sum_{i=1}^{N}({}_e\rho_i^q - \rho_i^t)^2 + E[\frac{1}{N}\sum_{i=1}^{N}({}_n\rho_i^q)^2] \quad (7.4)$$

Call the second term on the right ${}_nv^q$. Equation (7.4) shows that the expected mean-square error between test pattern and reconstruction is the sum of a deterministic part, which was examined in the noiseless experiments, and a part which may be measured from reconstructions of noise.

The pure noise was used with SIRT, GSIRT, and LSIRT2 (all unconstrained and unnormalized). In Figure 10 are plotted experimental estimates of $({}_nv^q)^{\frac{1}{2}}$ for the different algorithms as a function of q. The curves rise not quite linearly with q, SIRT rising at a rate only half that of the other two.

Qualitatively, then, we have an explanation for the phenomenon of Figure 9. The deterministic contribution to the distance diminishes as q increases; the random contribution increases. At some point there is a balance, and there we get, on the average, the best reconstruction. It would be desirable to stop the iteration at this point, but the position of this balance point varies with the algorithm, with the reconstructed object, and with the noisiness of the projection data. At present, there is no way of determining this ideal stopping point.

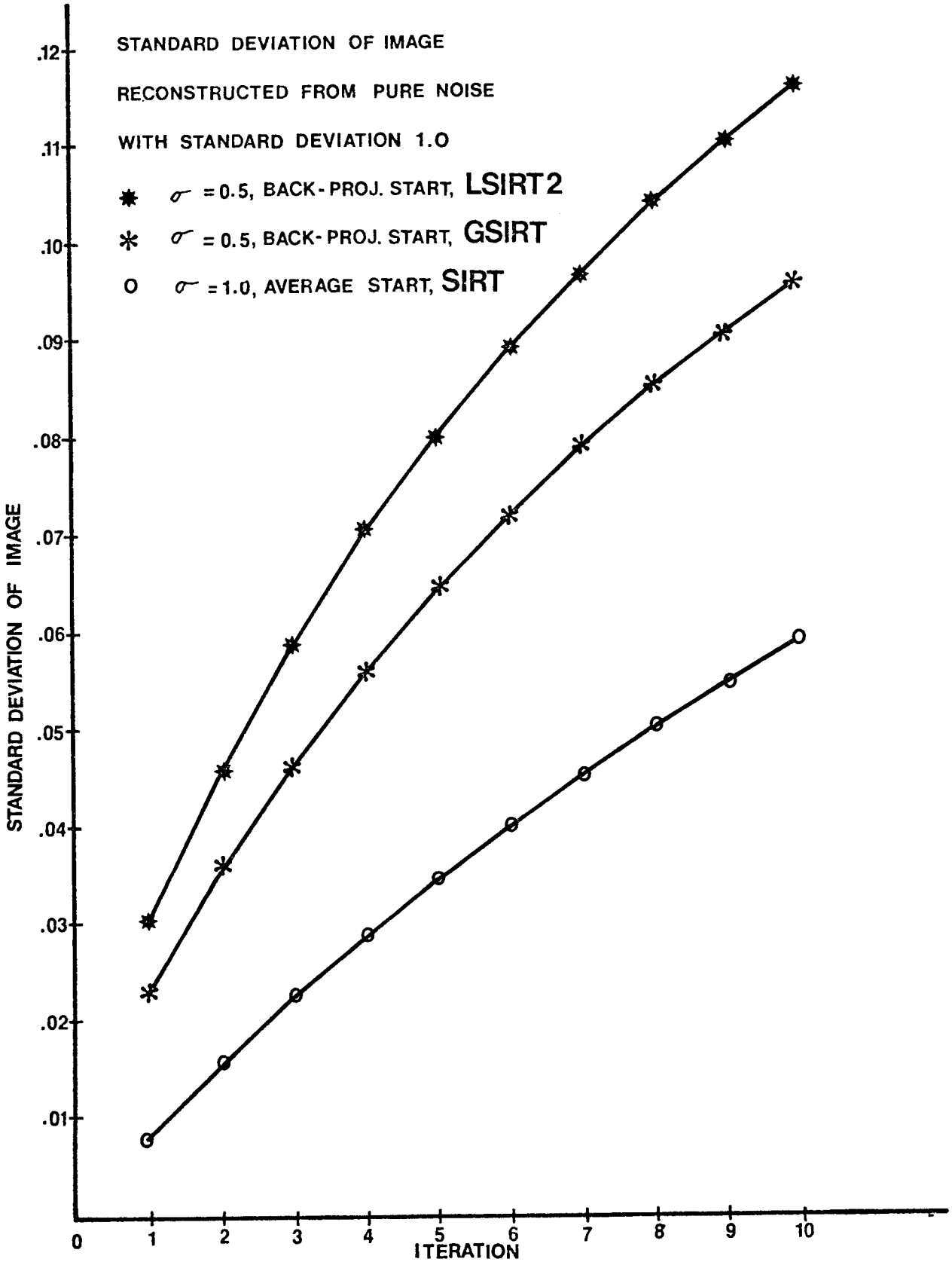


Figure 10. Reconstruction of 'pure' noise. A plot of the standard deviation of the reconstruction of pure noise versus iteration q.

● LSIRT2, ✱ GSIRT, ○ SIRT.

Note that SIRT amplifies noise only about half as fast as the true least-squares algorithms.

Conclusions

We have devised some new, faster algorithms which are modifications and extensions of SIRT. The GSIRT algorithm uses a back-projected picture of the Vainshtein type for the starting point and a relaxation factor for the correction term in the iteration scheme. The LSIRT1 and LSIRT2 algorithms are weighted least-squares extensions of SIRT which are computationally more economical than SIRT.

We have investigated experimentally the effect of noise in the projection data in the reconstructions produced by these algorithms. We have shown that SIRT, GSIRT, and LSIRT2 all amplify data noise as iteration progresses, SIRT at the slowest rate. However, the more rapid convergence of GSIRT and LSIRT2, as compared to SIRT, permits termination of these algorithms with fewer iterations, before the random contributions to the image reach a significant level.

ACKNOWLEDGEMENTS

The constant encouragement of Dr. Gabor Herman is gratefully acknowledged. Research for this paper is supported by NSF grant MCS75-22347, NCI contract CB53860, and NIH grants HL4664 and HL18968.

REFERENCES

1. A. V. Lakshminarayanan & A. Lent (1976), Proceedings of the SPIE/SPSE Conference on Application of Optical Instrumentation in Medicine V, Washington, D. C.

2. P. F. C. Gilbert (1972), J. Theor. Biol. 36, 105.

3. P. R. Halmos (1953), Finite-Dimensional Vector Spaces, Princeton: D. Van Nostrand.

4. A. Ben-Israel & T. N. E. Greville (1974), Generalized Inverses: Theory and Applications, New York: John Wiley and Sons.

5. G. T. Herman & S. W. Rowland (1973), Comput. Graphics & Image Process. 2, 151.

6. R. S. Varga (1962), Matrix Iterative Analysis. Englewood Cliffs, New Jersey: Prentice-Hall.

7. B. K. Vainshtein (1971), Sov. Phys. - Crystallogr. 15, 781.

8. R. A. Crowther & A. Klug (1974), Nature, 251, 490.

9. E. Whittaker & G. Robinson (1967), The Calculus of Observations, 4th edn. New York: Dover Publications.

10. M. Goitein (1972), Nucl. Instrum. Meth. 101, 509.

11. S. Rowland (1977), Computerized Reconstruction of Images from Projections, Based on Radon's Inversion Formula. Ph.D. Dissertation, Department of Computer Science, SUNY at Buffalo.

DISCUSSION:

Schmidlin started the discussion by noting that he and his colleagues at Heidelberg had investigated the SIRT algorithm, as well as a multiplicative ART algorithm. Their conclusion was that the ART algorithm gave much better results.

Lent agreed that there were probably better algorithms available than SIRT and noted that the multiplicative ART algorithm had the nice property that under some conditions it would give a maximum entrophy reconstruction.

Lent stated that the convergence to a least-squares solution should not be the only criterion for a proper reconstruction. He noted that very good least-square solutions often gave very bad visual reconstructions. Lent stated that the reconstruction program was so ill-conditioned that it required something beyond least-squares for a proper reconstruction. Metz asked a question concerning the comparison of the convergences of the various SIRT algorithms. In particular, the question related to the behavior of the algorithms for very large numbers of iterations. Lent stated that due to limitations on computer time, this region had not been explored. He noted, however, that the region that was explored extended well beyond what would be considered a good visual reconstruction. Further iterations tended to diverge from a good visual image.

Oppenheim asked how the proper number of interactions was determined. Lent indicated a visual criterion was probably best.

Weisner asked if one should put any dependence in the delta parameter of image quality. Lent noted that during low numbered iterations, delta behaved properly.

CONTRAST ENHANCEMENT OF SCINTIGRAMS BY THE COMPARISON WEDGES METHOD

I. Runczik, V. Černoch, B. Vavrejn and J. Kidery
Postgraduate Medical and Pharmaceutical Institute
and Institute for Clinical and Experimental Medicine
Prague, Czechoslovakia

ABSTRACT

The method of comparison wedges improves visualization of defects in high channel content gradient areas of the scintigram. The z - const. intersection planes commonly used for isocontour creation are replaced by linearly increasing channel content planes.

Three-dimensional objects are projected by the scintigraphic process into their two-dimensional images. These images are presented usually in a map form with information density data coding. The coding may be performed in various ways: shades of grey, colors, isocontours in computer based systems of display, etc. Equidistant isocontours are usually used. In the isometric display with a multichannel analyzer or computer the channel content is visualized by display matrix points shifted in the z-axis direction. The isocontour is obtained as a curve of intersection of anisometrically displayed plane and the constant z plane.

In our earlier paper[1] the difficulties of defect searching in areas with a high channel content gradient were demonstrated -(see Fig. 1)showing a normal scintigram of the liver in the upper left and equal defect simulated scinti-

grams in a three-color recycling display. The increasing channel content gradient is represented by a sequence of black, white and grey.

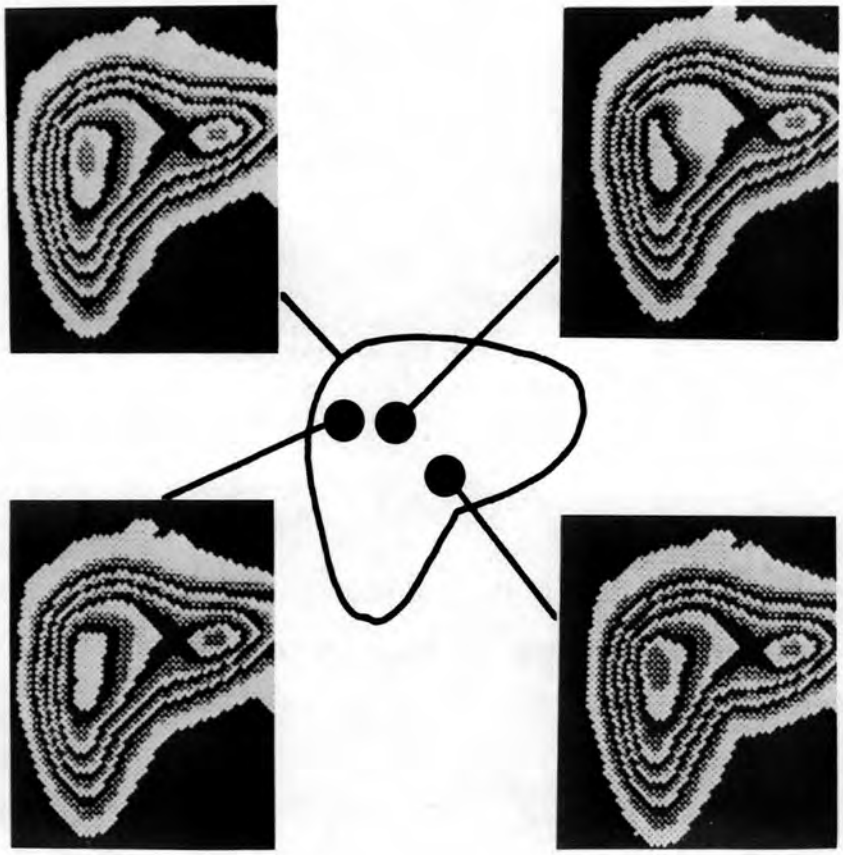


Fig. 1

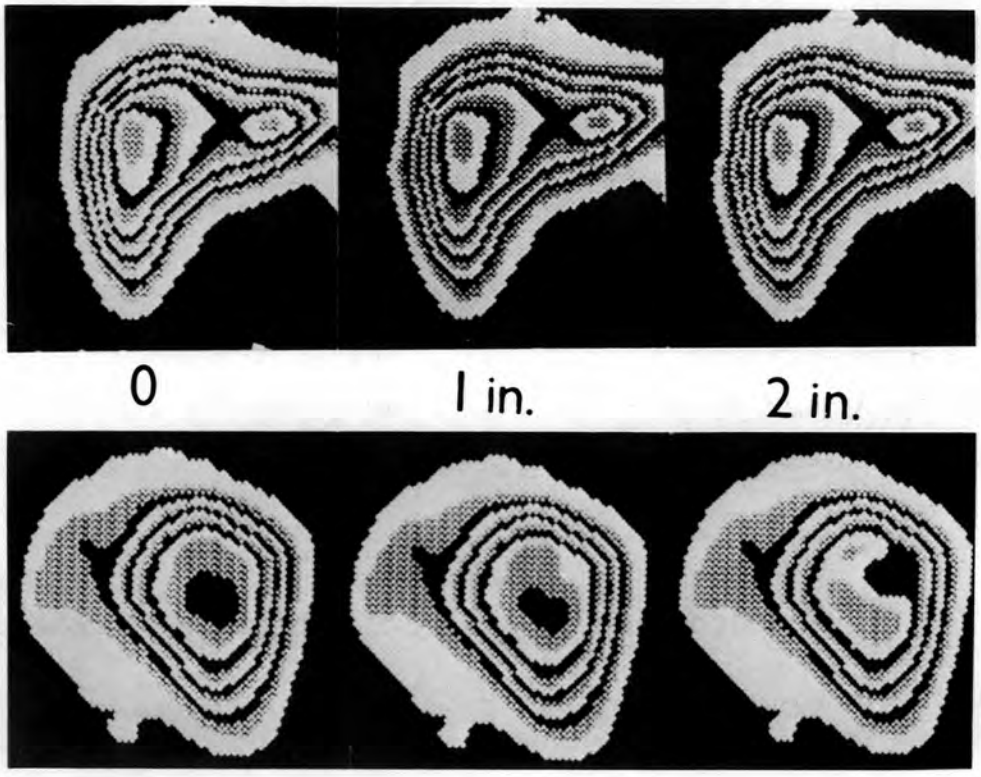


Fig. 2

As shown in Fig. 2, lateral projections help to overcome this problem. A-P projections are presented in the upper part and the lateral ones for the given defect diameters in the lower part of Fig. 2.

This paper deals with a new method of contrast enhancement of scintigrams in high gradient areas. This method is based on the use of comparison wedges. Fig. 3a shows standard intersection planes

$$z = \text{const.}$$

These planes are replaced in Fig. 3b by planes with a linearly increasing channel content, i.e.

$$z = c_1 x + c_2 y + c_0$$

Non-isocontours are obtained in the map form. From the point of view of defect finding, the non-isocontours can be, however, considered as valid as the commonly used isocontours.

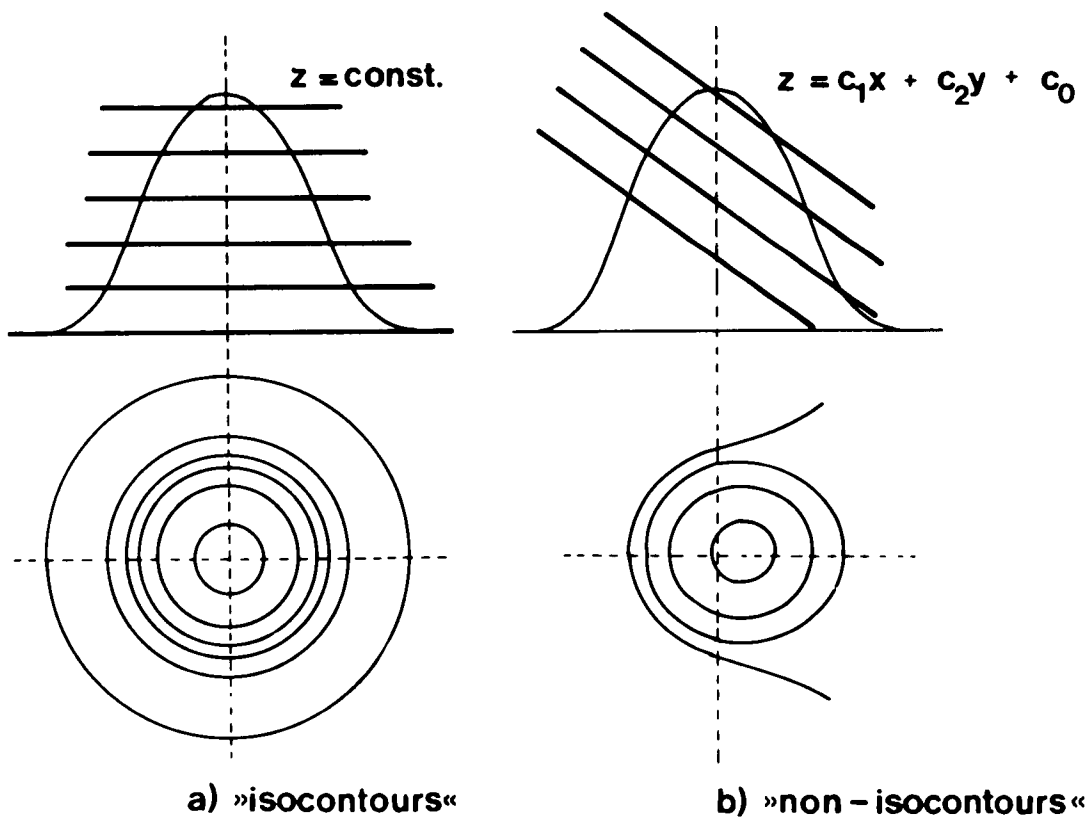


Fig. 3

The principle of this method is schematically represented in Fig. 4. The original high gradient profile is transferred to modified low gradient profiles by addition of linear wedges. To display the modified profiles in a standard way [isocontours with the z = const. intersection planes] a narrower dynamic range D is needed than for the original profile. Using the same density of the display matrix image forming points, the count rate differences can be visualized with an improved contrast.

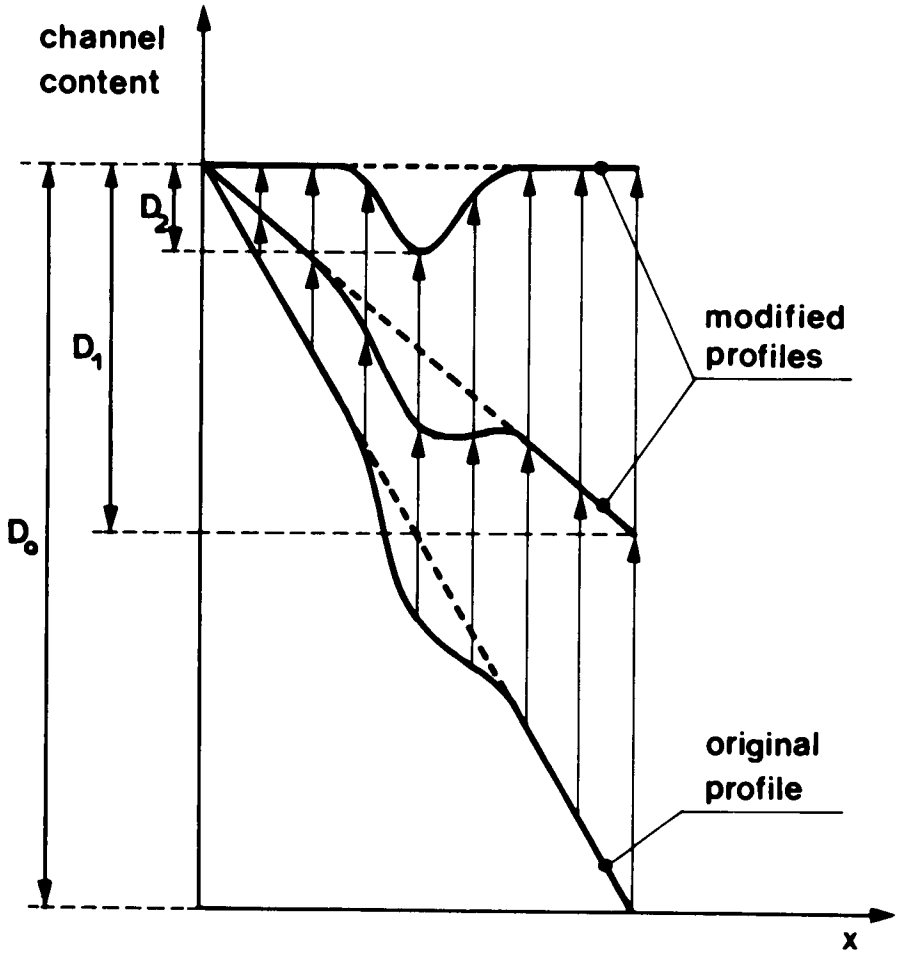


Fig. 4

We tested the comparison wedges method by means of a simple "roof-shaped" phantom with a defect on one declining area. In the random dot image [Fig. 5] of this phantom the defect is hardly visible.

For corresponding profiles and standard isocontour presentation see Fig. 6.

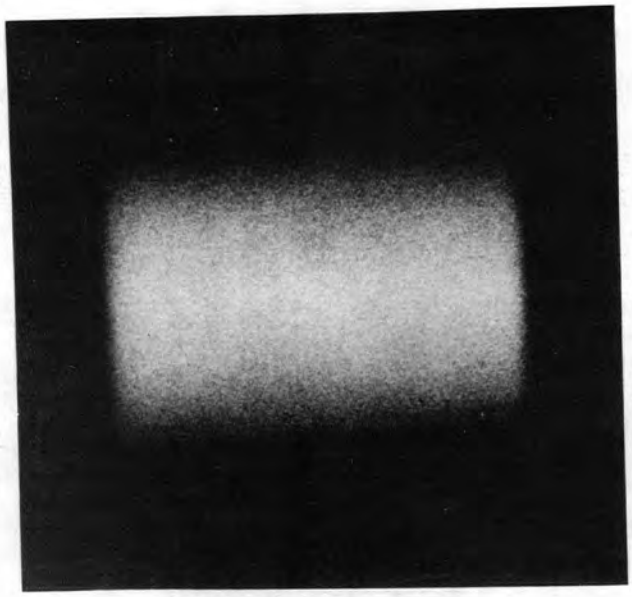


Fig. 5

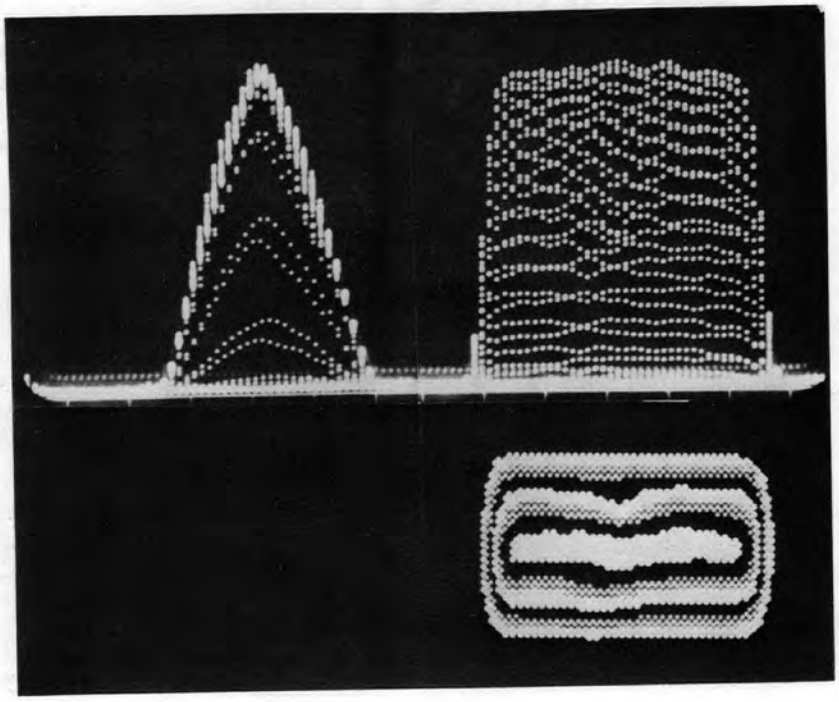


Fig. 6

Different slope wedges according to Fig. 7 were applied on the scintigram.

Fig. 8 shows the "roof-shaped" phantom depicted in the non-isocontour form by the comparison wedges method. Fig. 8a shows the original picture in isocontour presentation [corresponding to the zero slope wedge of Fig. 7a].

Figs. 8 b, c, d were obtained after the wedges shown in Figs. 7 b, c, d were applied to the "roof-shaped" phantom.

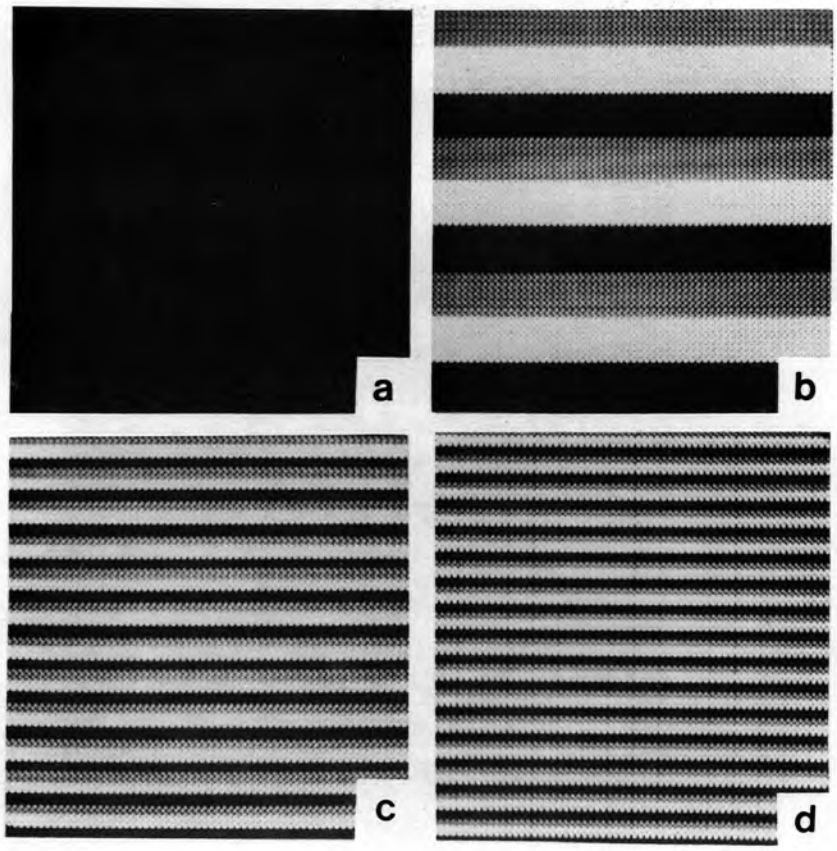


Fig. 7

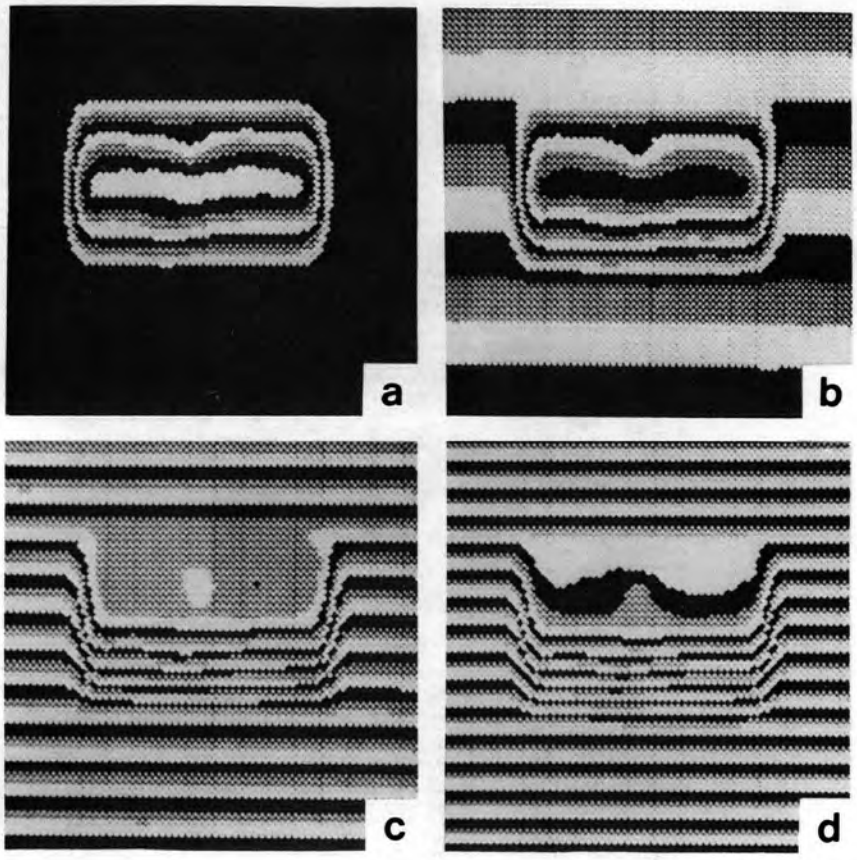


Fig. 8

The profiles related to Figs. 8a, c are given in Fig. 9. Note from Fig. 8c that we added such a slope wedge to obtain a minimum inclination of the modified profiles in the defect area. Modified profiles corresponding to Figs. 8b and 8d would be located in Fig. 9 below and above profile c, respectively.

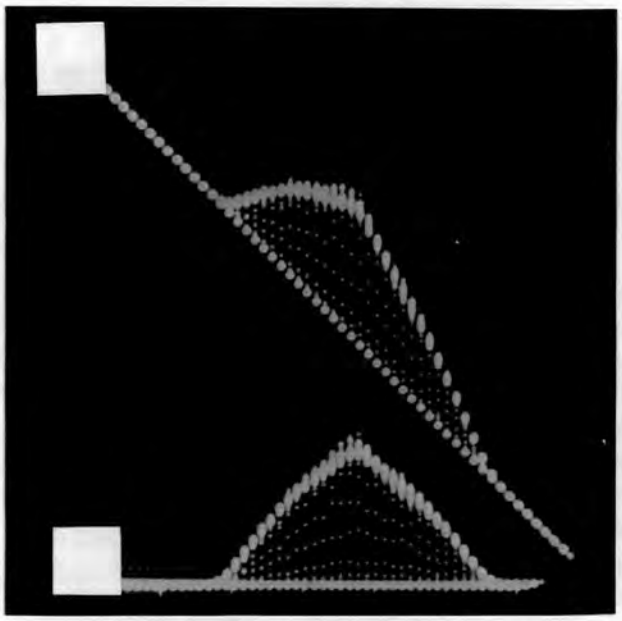


Fig. 9

Application of the comparison wedges reduced the dynamic range necessary for the display of resulting plateau thus permitting the increase of the number of display levels / Fig. 10 /.

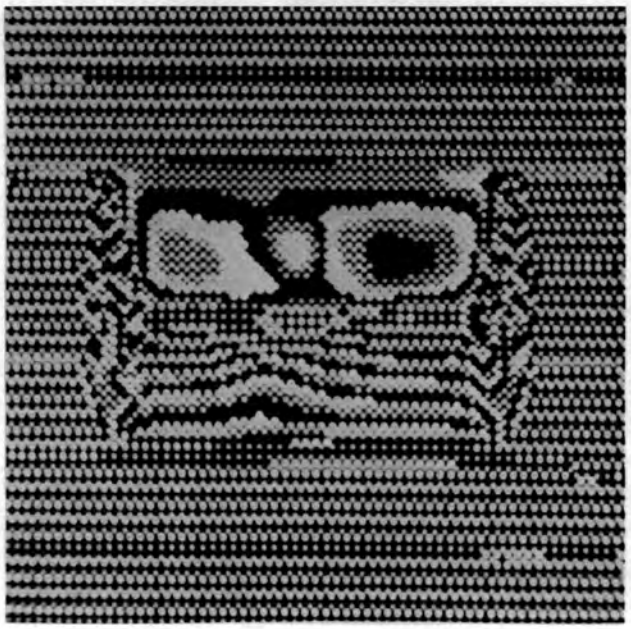


Fig. 10

Compared with the phantom study, the situation is more complicated when the described method is used for organ scintigrams. Due to considerable shape and gradient variability of individual areas it is useful to change the slope and orientation of comparison wedges. Examples of orthogonally and diagonally orientated wedges are shown in the quadrants of Figs. 11 and 12.

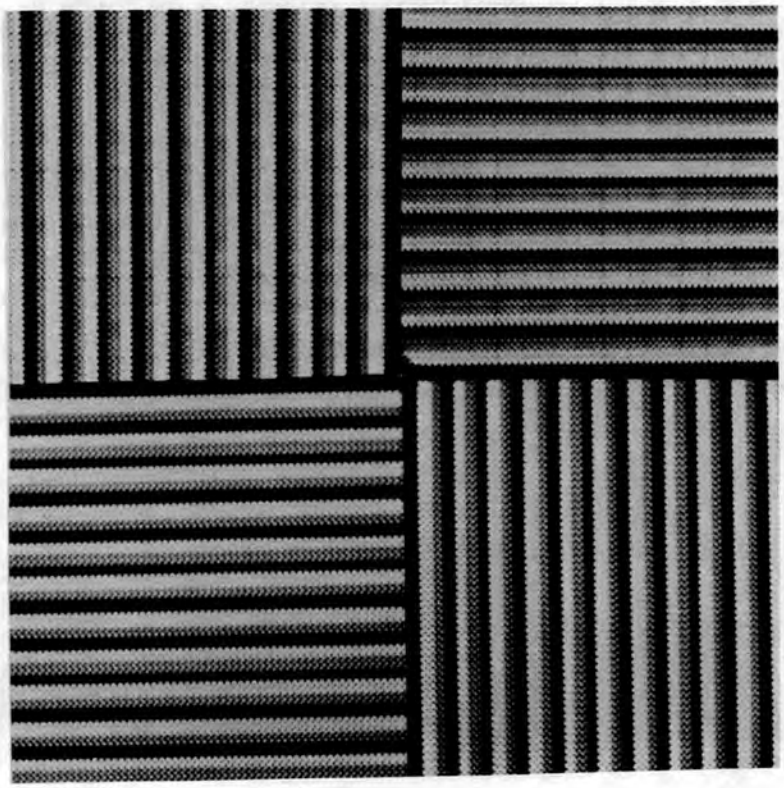


Fig. 11

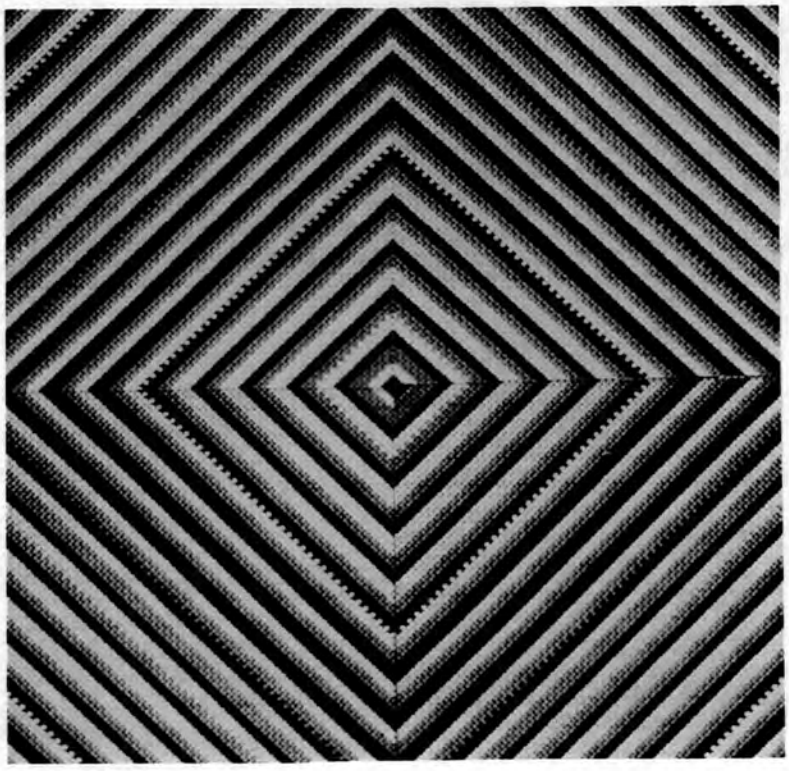


Fig. 12

These wedges are applied on a liver scintigram - see Fig. 13. The original scintigram is in the middle. In some areas of the scintigram there is a decreased gradient while in other areas we find an increased one.

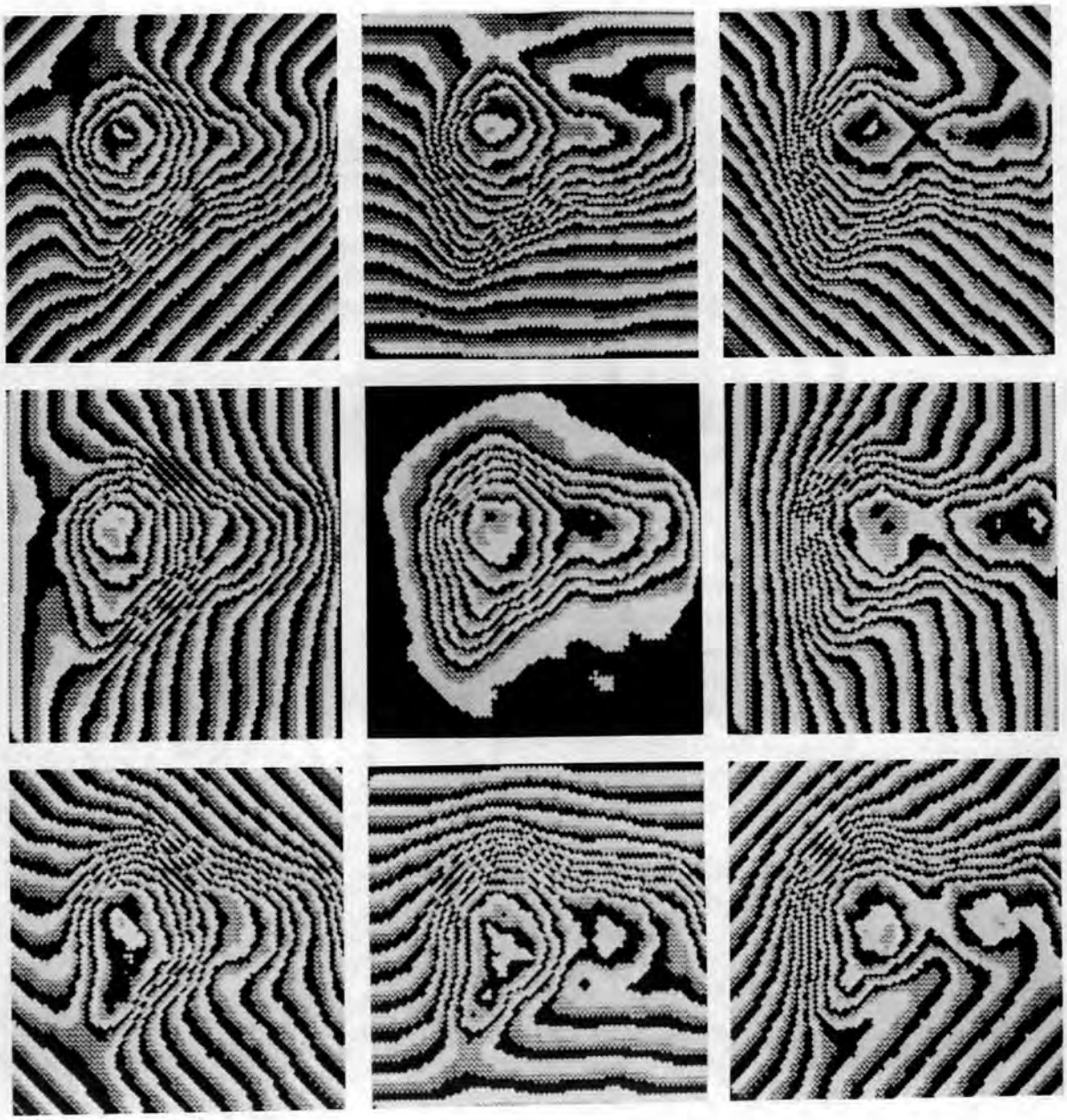


Fig. 13

Defect contrast enhancement by the comparison wedges method is illustrated by a study in Figs. 14, 15, 16. Fig. 14 shows the original liver scintigrams: Fig. 14a - without defect, Fig. 14b - with the defect simulated in the border area of the right liver lobe. This defect can hardly even be found. The basic comparison wedge and its multiplies were applied to both scintigrams.

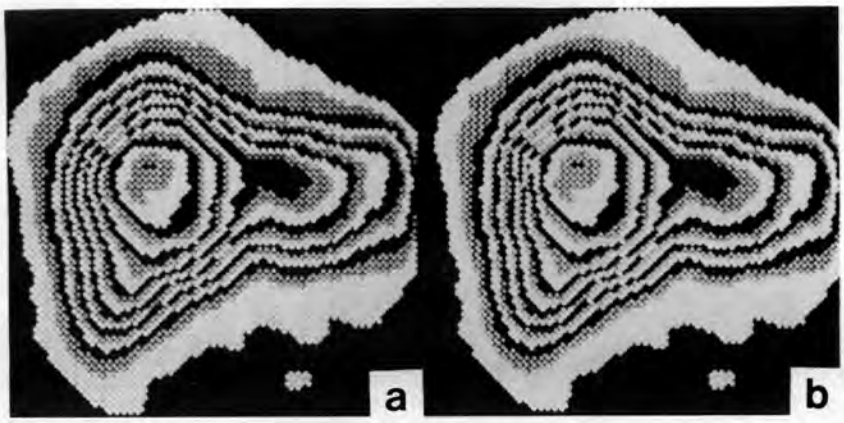


Fig. 14

Fig. 15 shows the resultant images in the non-isocontour presentation for the scintigram without defect.

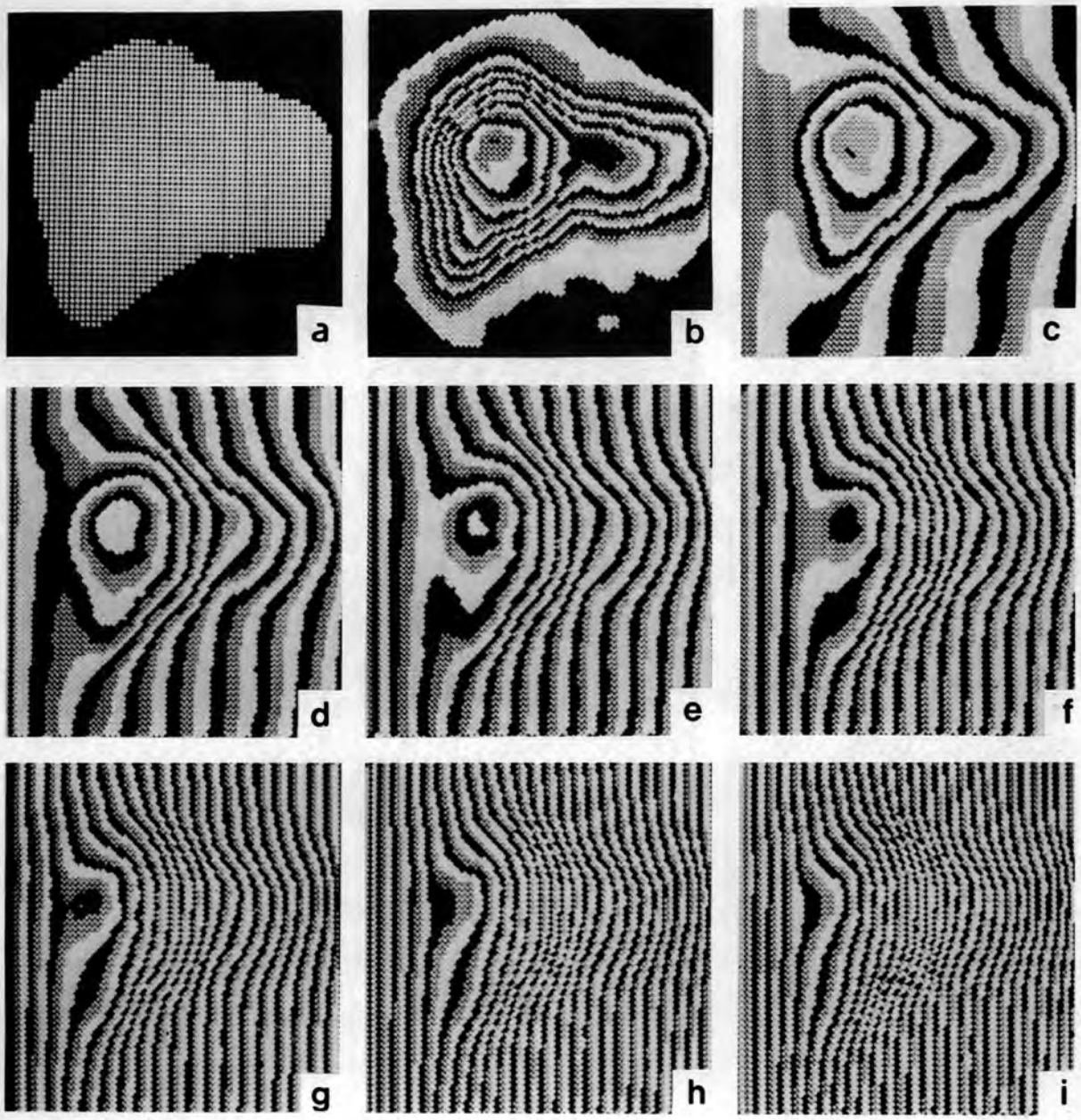


Fig. 15

Corresponding images are given in Fig. 16. Localization of the simulated defect is shown in Fig. 16a. The defect is visualized with a maximal contrast in Figs. 16e - g.

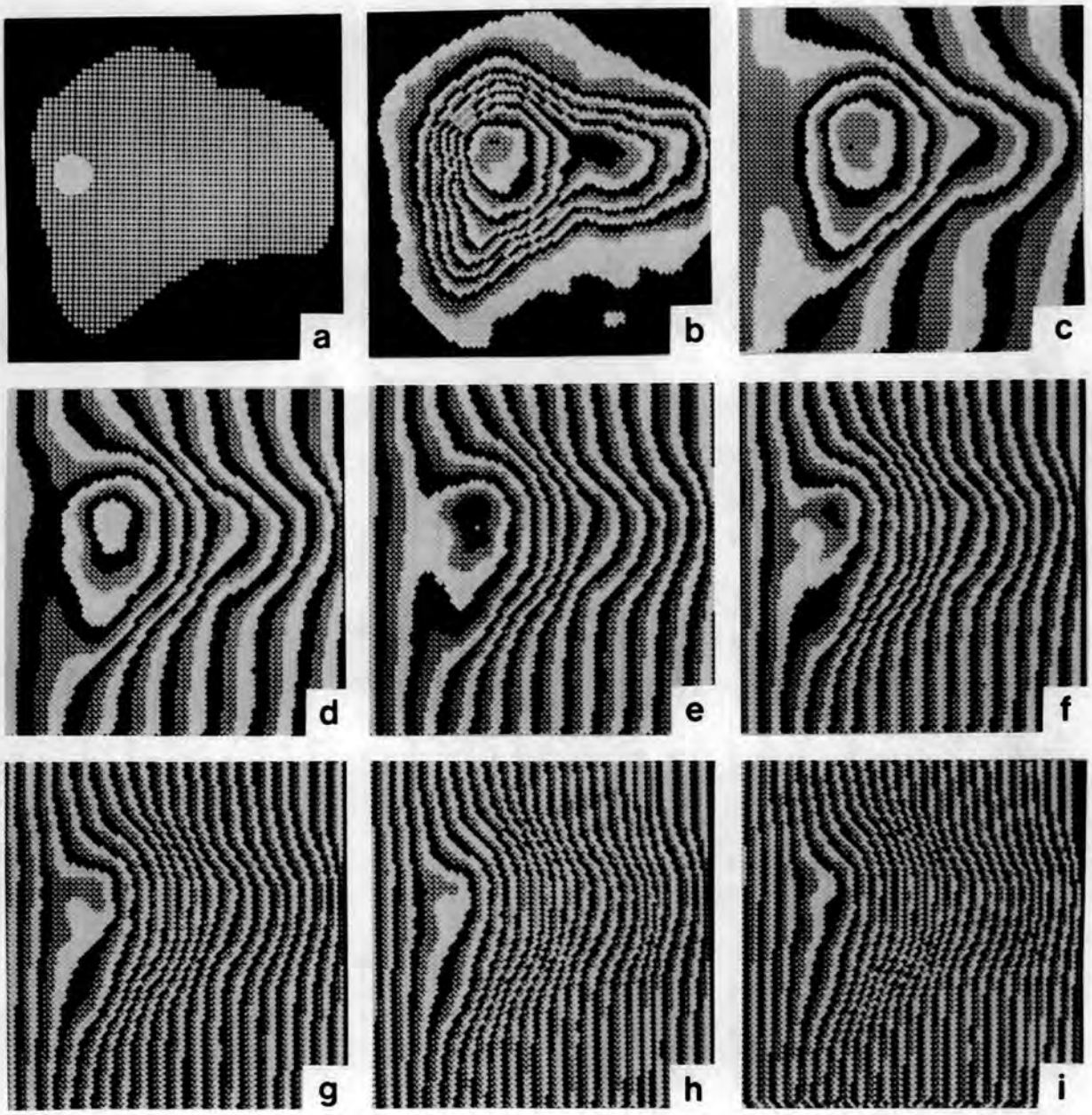


Fig. 16

CONCLUSION

The described method of the comparison wedges is aimed at contrast enhancement in high channel content gradient areas of scintigraphic images and thereby better defect visualization.

REFERENCE

1. I. Runczik, V. Černoch, B. Vavrejn, "Hybrid simulation: a new method for comparison of scintigraphic devices" p. 691 in Medical Radioisotope Scintigraphy 1972 [Proc. Symp. Monte Carlo, 1972], Vol. 1, IAEA Vienna, 1973.

DISCUSSION:

Ochs noted that similar images would be obtained if one subtracts a very heavily smoothed image from the original. Runczik agreed but noted that there would be a problem of "dot shifting" as a result.

Tweig asked if techniques of comparison wedges could be used as well with displays other than isometric contour displays. Runczik noted that the technique was not dependent upon the type of display used. Runczik also noted that the statistical integrity of the image was left unchanged after the addition of the "noise-free" wedge.

Metz asked if it was possible to get crossing contours if the wedge were too steep. Dr. Runczik stated that the technique did not involve a rotation but only an addition and would, therefore, not cause a crossing of contours.

COMPARISON OF SOME NONLINEAR SMOOTHING METHODS*

P. R. Bell and R. S. Dillon
Health and Safety Research Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee 37830

ABSTRACT

Due to the poor quality of many nuclear medicine images, computer-driven smoothing procedures are frequently employed to enhance the diagnostic utility of these images. While linear methods were first tried, it was discovered that nonlinear techniques produced superior smoothing with little detail suppression. We have compared four methods: Gaussian smoothing (linear), two-dimensional least-squares smoothing (linear), two-dimensional least-squares bounding (nonlinear), and two-dimensional median smoothing (nonlinear). The two-dimensional least-squares procedures have yielded the most satisfactorily enhanced images, with the median smoothers providing quite good images, even in the presence of widely aberrant points.

INTRODUCTION

Smoothing procedures are frequently used on nuclear medicine images to improve readability or to permit use of color displays. Simple smoothing methods such as the nine-point or a Gaussian weighting matrix characteristically cause a rather rapid flattening of sharp image detail. Such methods are not good bases for nonlinear bounding because of this detail flattening. The smoothing method using a least-squares fit to a two-dimensional polynomial[1,2] does not produce rapid detail flattening and is therefore a suitable method to use as a basis for a bounding procedure.

*Research sponsored by the Department of Energy under contract with Union Carbide Corporation.

By acceptance of this article, the publisher or recipient acknowledges the U.S. Government's right to retain a nonexclusive, royalty-free license in and to any copyright covering the article.

The detail flattening performance of two 7x7 Gaussian weighting matrices and a 7x7 least-squares smoother are compared in Figure 1. A 64x64 element Poisson noise field was generated for N=5. This field contained a 3x21 element

ORNL DWG. 77-10962

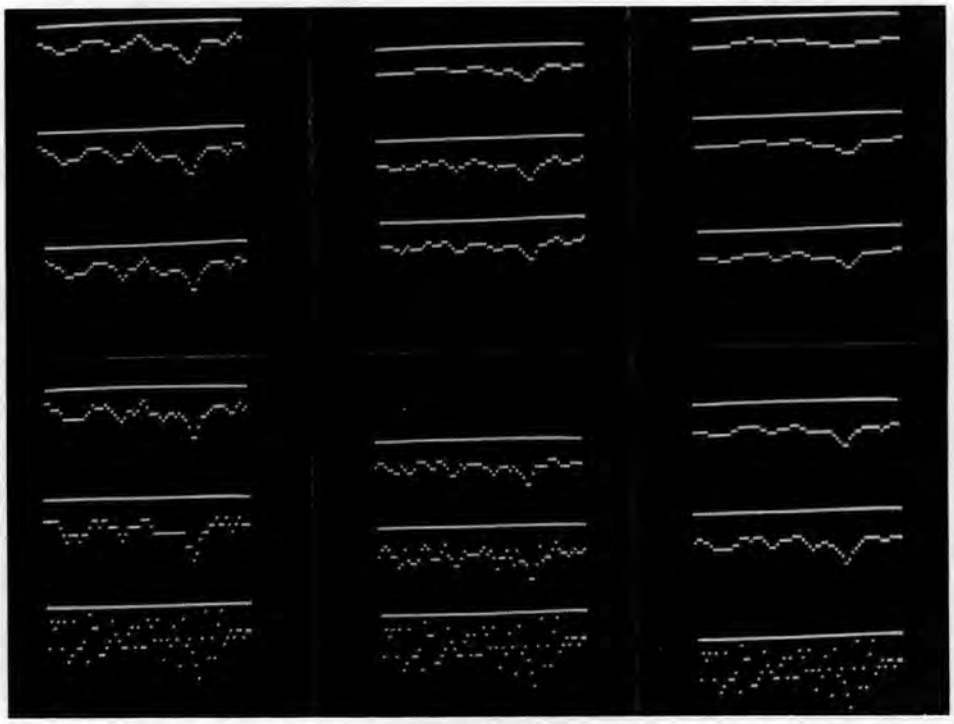


Fig. 1. The results of three smoothing methods applied to a 64x64-element Poisson field (N=5) with an embedded 3x21-element signal area (N=8). Shown are the profile views along a line perpendicular to the higher signal area at its center. In all columns the top line is that of the raw data. The second line shows the results of using the matrix as a bounder at .7 standard deviation. The following lines are the results of successive passes as a smoother; the last line had four extra passes. At left a wide Gaussian (0.7 standard deviation per picture element) was used. The center column was produced using a narrower Gaussian (1.0 standard deviation per picture element). The right column was produced using a 7x7-element least-squares fit to a two-dimensional quartic.

signal area of N=8. This represents a 45 counts/cm^2 background area with a 72 counts/cm^2 signal bar (or image "detail") 1x7 cm in size. The image scale corresponds to approximately three camera elements per centimeter. The intensity profiles of Figure 1 were taken perpendicularly to the long axis of the signal bar. In each column the top line is the raw data line and the lines below have successive processing passes. The bottom line has four additional passes making a total of eight. A Gaussian matrix with 0.71 standard deviation (SD) per element was used for the left column, a sharper Gaussian with 1.0 SD/element for the center and a 7x7 element two-dimensional quartic for the right column. The first pass of the quartic was a bounder bounding at 0.7 standard deviation from the fitted value (see description of bounder presented below). The unique elements of each matrix are listed in Table 1. After two passes of each process, the 1-cm wide signal widths at half maximum were 9, 6, and 4 elements, respectively.

Table 1. Unique Elements of Smoothing Matrices

LOCATION	WIDE GAUSSIAN	NARROW GAUSSIAN	LEAST SQUARES QUARTIC
0,0	0.1595	0.3155	0.1779
1,0	0.0966	0.1176	0.1342
1,1	0.0588	0.0433	0.0928
2,0	0.0215	0.0059	0.0421
2,1	0.0129	0.0022	0.0074
2,2	0.0029	0.0001	-0.0575
3,0	0.0017	0	0.0184
3,1	0.0011	0	-0.0050
3,2	0.0003	0	-0.0359
3,3	0.0001	0	0.0425

The two-dimensional least-squares polynomial smoothing method resembles to some extent the familiar orthogonal polynomial method in one-dimension. It requires 49 address calculations and 10 multiplications for each data point but produces good smoothing and has the desirable property that the weight matrices may be added together to produce combined processes in one pass. Figure 2 shows the application to a liver phantom with a 17-ml void. The image contains 135,000 counts. The void can be seen in the left image but is more easily found in the center multicycle image. By processing with a matrix made by adding the matrix for the 5x5 quadratic smoother to -0.9 times the 7x7 flat smoother, the slow changes in the image are suppressed while the more rapid changes are retained. Application of this filtering matrix causes the void to be very obvious as seen in the right image. This procedure resembles the unsharp smoothing method (Canterbury filter).

ORNL DWG. 77-10961

Fig. 2. The two-dimensional least-squares method applied to a liver phantom with a 17 ml void. At left is the normal image map; the multicycle display is shown in the center. At right is the result of the image processed by one pass of the sum of the 5x5-element quadratic and -0.9 times the 7x7-element flat matrix, showing a reduction in slowly varying image components.

Nonlinear bounding has a considerable effect in reducing noise "blobs" in an image as demonstrated in Figure 3, where a noise field similar to that for Figure 1 was processed by the nine-point smoother (left) and the nonlinear method (right). The noise "blob" height is reduced about a factor of two while retaining about the same relief of the image detail.

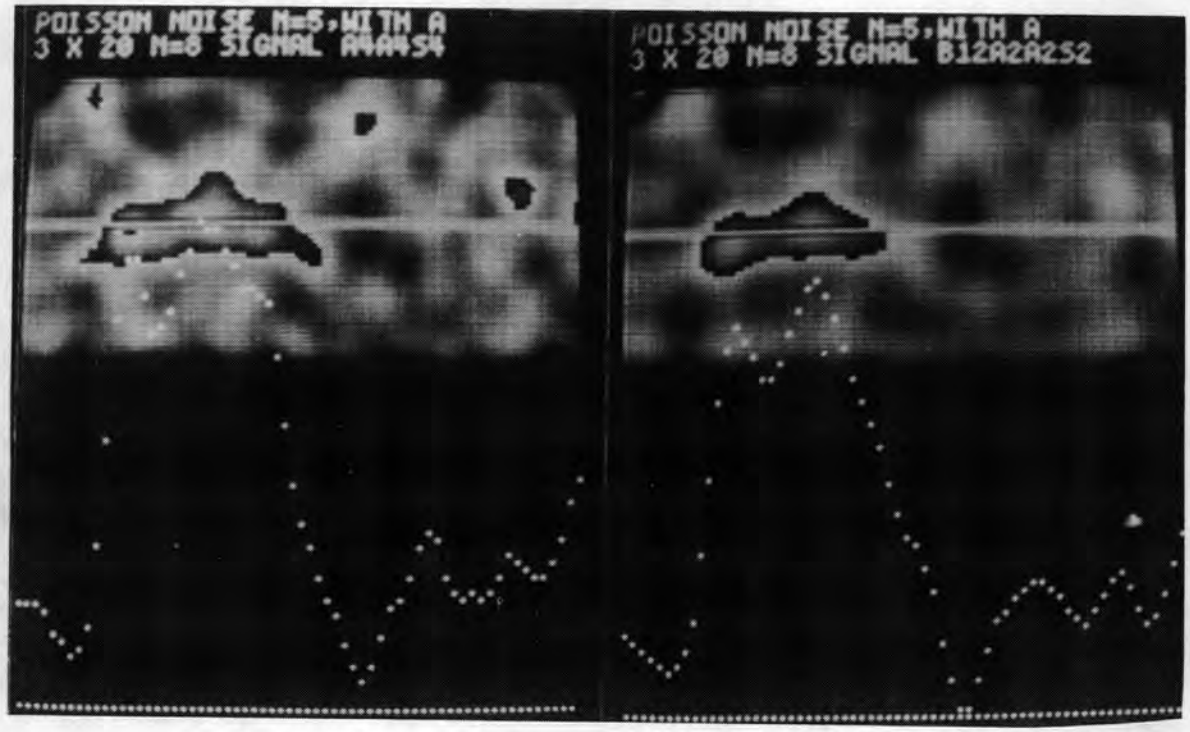


Fig. 3. A Poisson noise field (N=5) with an embedded 3x21-element signal area (N=8) processed by the nine-point smoother (left) and by the nonlinear least-squares method (right). The noise fluctuations (blobs) are reduced by the least-squares method without reduction of the more continuous signal detail.

The "nonlinearity" of the process is embodied in the bounding procedure in which the fitted value at the center of the 7x7 area is compared with the data point. If the difference between the values is more than (in this case) 0.7 times the square root of the fitted value, the data point is replaced by the fit; if the difference is smaller, the data point is not replaced. This decision affecting the result is the essential process making the nonlinearity. In this same sense the process described by Gustafsson and Pizer[3] is a nonlinear one, because data points either are or are not included in the process, depending upon their effect on the result. The older process[4] of nonstationary smoothing, which used a larger number of points for smoothing in areas of lower counting rate, was not nonlinear since the individual points were included without any examination of their values.

An effective one-dimensional nonlinear smoothing process depending upon the use of the median of a consecutive group of data points was developed by Tukey.[5,6,7] In this process the median of 3 or 5 data points is used to replace the central point of the range for every such range in a data line. The resulting line of data is smoothed by some suitable process and the smoothed line is subtracted from the original data. The difference line is similarly treated, and the result is added to the first smoothed line as a correction. This double processing method was called "twicing" by Tukey. We have made a simple extension of this method to two-dimensions using a 3x3 or 5x5 square area and using the least-squares matrix smoother in the smoothing step. The process is exceedingly tolerant of points in gross error as they are ignored completely, whereas the other processes need special protection against points ten or more times larger than the mean in its neighborhood.

A comparison was made between the median and the least-squares smoothers on a ^{67}Ga scan of an abdomen. The center image of Figure 4 was produced by "twicing" with the 5x5 element median method, the right image by the least-squares bounder. The lower detail relief of the median method is evident.

ORNL DWG. 77-11571

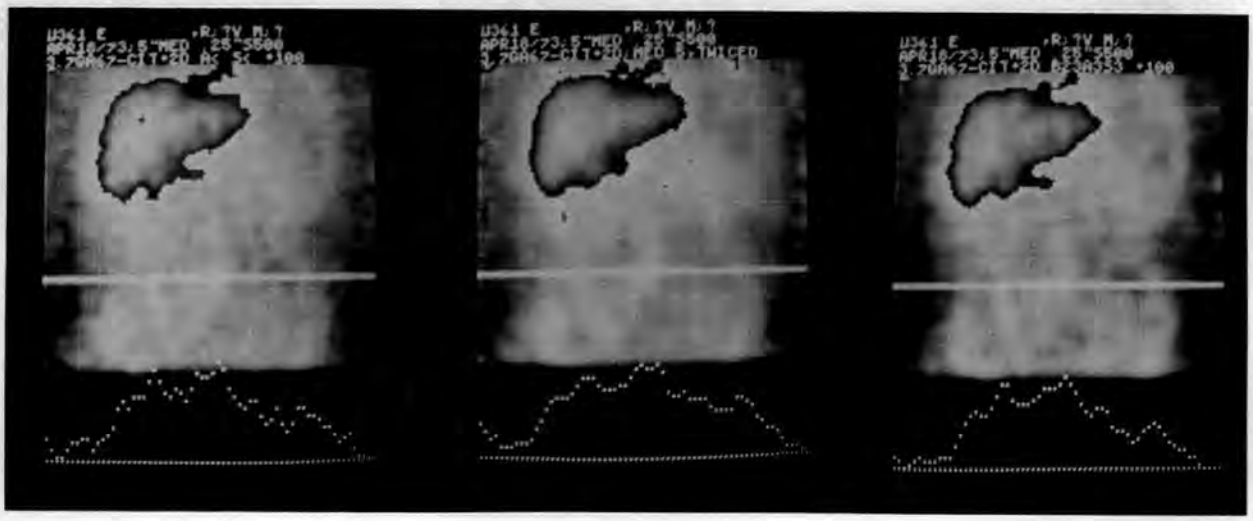


Fig. 4. A ^{67}Ga image of an abdomen processed by the 7x7-element quartic (left), the 5x5-element median filter (center) and the two-dimensional least-squares method (right). The intensity versus position plot at the bottom of each display was made at the location marked by the bright line in the image.

Both the median method and the least-squares methods would be more satisfactory for use on dynamic sequences where they are most needed if they could be made more rapid. One-dimensional methods applied in two orthogonal directions to an image require less time but produce poorer results. The seven-point least-squares method applied in X and Y directions requires 14 address calculations and eight multiplications per data point. The median method applied in X and Y is also faster than in two-dimensions although it is always

slower than the least-squares method. The images (of the same subject as in Figure 4) in Figure 5 are: left, two boundings followed by two smoothing passes in the X direction followed by the same in the Y direction using a seven-point quartic; center, the X and Y processes of each type were alternated; right, two-dimensional processes. The all X then all Y image is quite ragged. The alternate X and Y image shows good detail relief, but the continuity of the image detail is less satisfactory than in the two-dimensional method. See, for example, the ascending colon (lower left of center).

ORNL PHOTO 2739-77

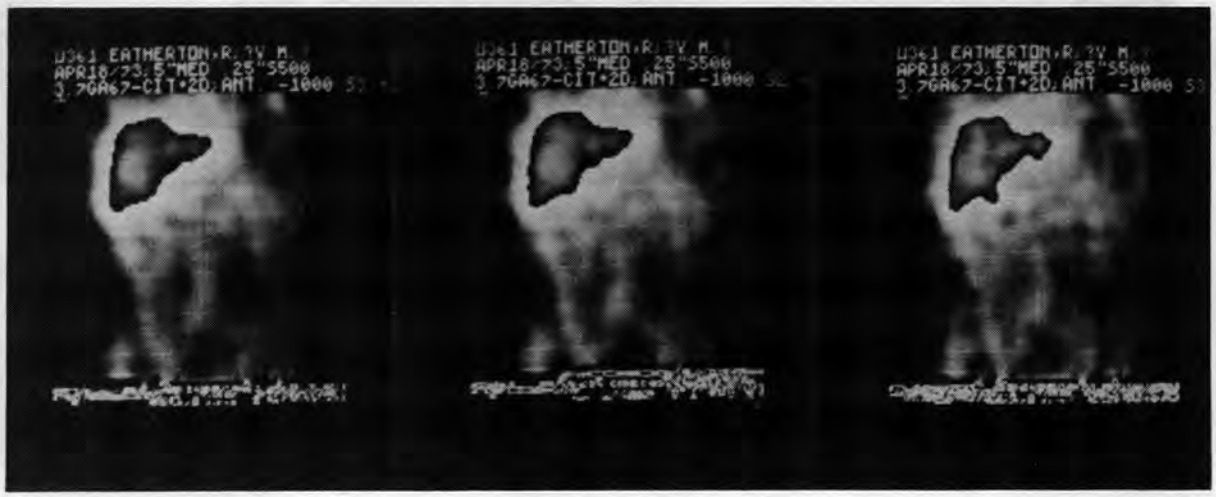


Fig. 5. A comparison of repeated X and Y one-dimensional bounding and smoothing and the two-dimensional applications of the least-squares method. At left all X processes preceded all Y processes. At center the X and Y processes were alternated. The results of the two-dimensional method are shown at right.

The median method shown in Figure 6 (applied all in X followed by all in Y at the left, alternate X and Y in the center and two-dimensional at the right) show less differences. The detail is less blurred in the two-dimensional case.

The methods were then applied to the liver phantom image that had much higher data values. Figure 7 shows the alternate X and Y applications of the seven-point quartic on the left and the 7x7 two-dimensional quartic on the right. The alternate application is too ragged and thus unsatisfactory, and the two-dimensional quartic is barely smooth enough for detection of the void (probably the quadratic should be used for such a smooth image). Note that the dip due to the 17-ml void is about 15%. Figure 8 shows the application of the alternate X and Y five-point median on the left and the 5x5 two-dimensional median on the right. The differences are less striking but the two-dimensional method is smoother.

ORNL PHOTO 2736-77

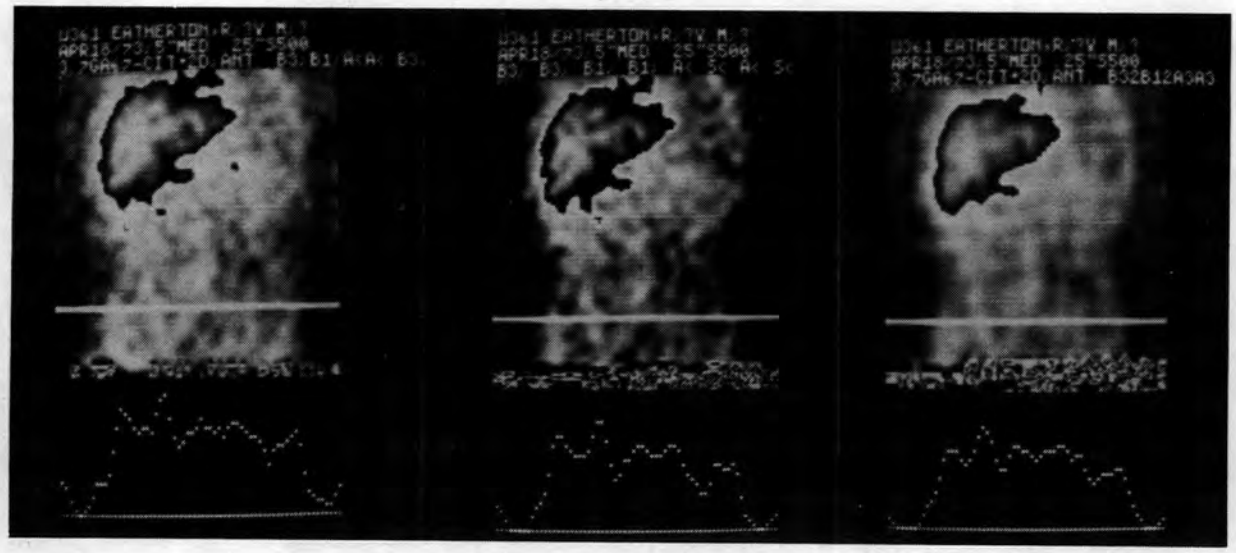


Fig. 6. Comparison of separate X and Y one-dimensional bounding and smoothing and the two-dimensional applications of the median method. At left all X processes preceded all Y processes. At center the X and Y processes were alternated. The results of the two-dimensional method are shown at right.

ORNL PHOTO 2738-77

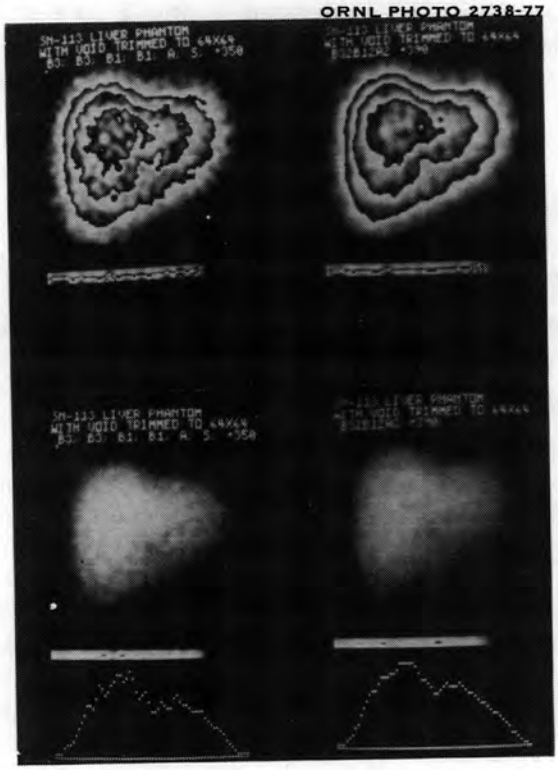


Fig. 7. The liver phantom processed by alternate X and Y applications of a one-dimensional seven-element quartic (left) and by the two-dimensional quartic method (right). The top image in each column is the multicycle display of same data shown in normal map below.

ORNL PHOTO 2737-77

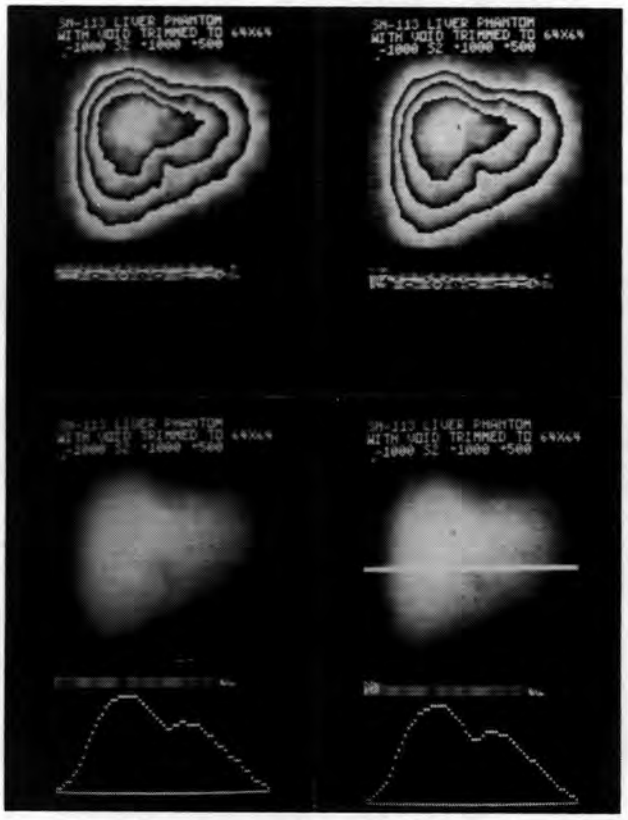


Fig. 8. Alternate application of a one-dimensional five-element median filter and a seven-element quartic in X and Y (left) and the results of the matching two-dimensional process (right).

CONCLUSION

Two nonlinear smoothing methods were compared: the two-dimensional least-squares polynomial method and a two-dimensional version of the Tukey median method. The median method provides remarkable tolerance for very bad points but does not give as good detail relief as the two-dimensional least-squares polynomial method. Two orthogonal one-dimensional applications of the median method are only slightly poorer in image quality. The least-squares method is faster and gives the best detail relief, but successive orthogonal one-dimensional applications of this method are quite unsatisfactory.

REFERENCES

1. M. R. Bell, A One- and Two-Dimensional Least-Squares Smoothing and Edge-Sharpening Method for Image Processing, ORNL/TM-5222 (January 1976).

2. P. R. Bell, R. S. Dillon and M. R. Bell, "Nonlinear Image Processing," pp. 392-400 in Proceedings of the Sixth Symposium on Sharing of Computer Programs and Technology in Nuclear Medicine, Atlanta, Georgia (January 25-26, 1976).

3. T. R. Gustafsson and S. M. Pizer, "Non-Stationary Metz-Filtering," pp. 56-65 in Proceedings of the Fourth International Conference on Information Processing in Scientigraphy, Osay, France (July 15-19, 1975).

4. S. M. Pizer and H. G. Vetter, "Processing Radioisotope Scans," Journal of Nuclear Medicine, 10, 150-154 (1969).

5. J. W. Tukey, "Nonlinear (nonsuperposable) Methods for Smoothing Data," p. 673 in The Congressional Record, EASCON (1974).

6. A. E. Beaton and J. W. Tukey, "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band Spectroscopic Data," p. 147 in Technometrics, 16 (1974).

7. L. R. Rabiner, M. R. Sambur and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23, 6 (December 1975).

THE IMPLEMENTATION OF A FLOPPY DISC DATA LOGGER FOR MEDICAL IMAGES:

A PRACTICAL APPROACH TO THE REPORTING STATION CONCEPT

N.J.G. Brown, D.F.R. Burns and K.E. Britton
Departments of Nuclear Medicine, The Middlesex Hospital Medical School
and St. Bartholomew's Hospital, London, U.K.

ABSTRACT

The concept of a central reporting station implies a data logger on each data collection device, X-ray, Thermography etc. and a common data analysis and display system capable of meeting the requirements of each image type. A double density floppy disc system which is at present under construction, meets the data logging requirement for all but X-ray images where a digitising system described earlier in this meeting is effective. This concept makes for more efficient use of expensive data processing facilities and allows the interaction of different diagnostic methods from which clinical strategies can be developed. The processing system at The Middlesex Hospital developed originally for nuclear medicine has so far been used also for radiographs and thermograms. Further development of the concept of the reporting station is for discussion

---

THE REPORTING STATION

The fundamental principle of the reporting station concept is that each imaging device (gamma camera, scanner, CT scanner etc.) has a data logger attached to it. The data analysis and display system, which we call the reporting station, is able to handle and display data from any imaging device. Ideally only one or at most two forms of recording medium are used for all data. Several advantages come from this scheme; the processor, with its expensive display systems, is more efficiently used when it is shared by a number of imaging devices- the time of the technologist and the imaging equipment itself is also better utilised when the latter is housed separately from the processor; simultaneous processing and data collection from as many devices as desired can go on without conflict. Furthermore the imaging devices need not be in the same building as the processor. Indeed they may be in a different institution some miles away. The reporting station, however, can be in the place which is most convenient for those who do the reporting, and if located in the radiologists' reporting room it could act as a catalyst in the integration of the results obtained by different modalities. Finally, a further factor which makes for greater efficiency by comparison with on-line systems is that data need only be entered into the processor if it is to be analysed.

THE RECORDING MEDIUM

A satisfactory short term recording medium for use with the reporting station should be convenient, reasonably priced, and be read by devices which can be easily interfaced to existing image processing systems so that they may be used as reporting stations. It must have adequate capacity and recording rate.

All the above requirements appear to be met by a double density floppy disc readable by a drive with commercially available interfaces for the DEC, Varian and Data General computers on which Nuclear Medicine image processing systems are based. A double density disc has 77 tracks and 32 sectors each containing 256 8-bit bytes. Thus since each track can hold two 64 x 64 frames at 8-bit accuracy a whole disc can, for example, accommodate 154 frames of 64 x 64 images or 38 frames of 128 x 128 images or 9 frames of 256 x 256 images at 8-bit accuracy. Only half these numbers of frames can of course be stored at 16 bit accuracy, for example.

All devices operate at one revolution in 166 mSec. A particular drive marketed by Sykes features a track to track time of 6 mSec and a head settling time of 43 mSec. Thus the time required to dump one track is approximately 380 mSec (assuming two revolutions are required to write the data) allowing five 64 x 64 bit frames to be dumped each second.

This performance is considered sufficient for all the data logging needs of the average Nuclear Medicine department. It would be convenient to use one disc per patient for dynamics and one disc per machine per session for statics. At report time the disc data can be conveniently loaded into the processor as required.

IMPLEMENTATION

A floppy disc data logger for use with a gamma camera is currently under construction at The Middlesex. A block diagram of the system is shown in fig. 1. A special image memory, which may be configured in different ways, is connected to an interface and control unit. The latter receives data from dual analogue to digital converters and acts as an interface between the image memory and the memory bus of an 8K Micro Nova computer which controls the system. By setting a control register in the interface and control unit the computer is able to set up the memory configuration and the mode of data collection (matrix size, single or dual isotope). The interface unit contains the necessary logic to increment the appropriate memory location when a value is presented to it by the ADCs. It is so designed that the parts of the image memory which are allocated to the micro Nova appear to it as part of its own memory. The whole image memory may be allocated to the ADCs or computer. Alternatively half may be allocated to the ADCs while the other half is allocated to the computer. Thus allowing data collection and dumping to disc to proceed concurrently.

The double density drive is used for both program entry and data logging. The system is set up for a particular task by simply loading the appropriate disc. The switch indicator panel is used to input date and time at the beginning of the day and patient identification numbers. Camera run/stop status is available to the computer so that it can start and stop storage at the right time without the need of attention by the camera operator.

CONCLUSION

The system described provides a flexible test bed for remote data logging and the reporting station concept in Nuclear Medicine. Future modification and development will allow it to deal with data produced by other modalities.

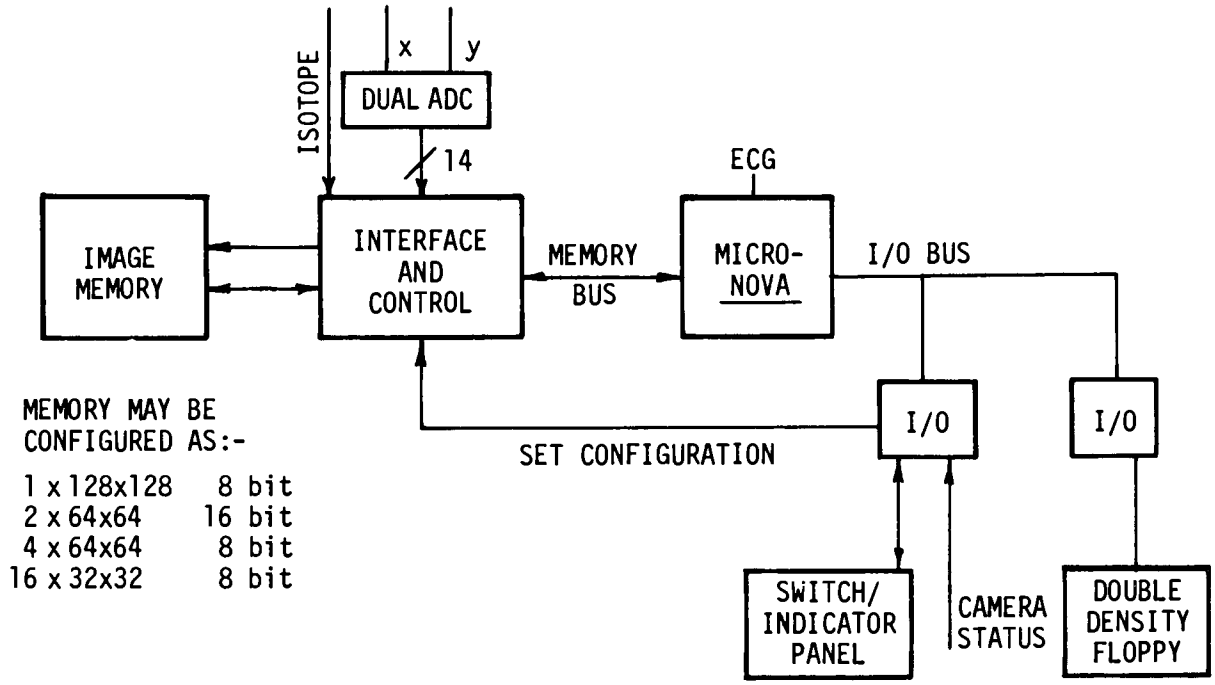


Fig. 1. Block Diagram of System

DISCUSSION:

McClain asked a question concerning the reliability of the floppy disk recording surface. Brown responded by noting that there was a floppy in his institution which had operated error free for the past nine months.

Goldberg commented that they had experienced very good reliability with their floppies. He also made an appeal that a standard data format for floppies be adopted. In addition, he noted that they had been able to achieve data collection rates of 30 (32 x 32) image/second with their floppy.

FLUOROSCOPIC SUBTRACTION IMAGING VIA REAL-TIME DIGITAL VIDEO PROCESSING--CURRENT WORK**

Theodore L. Houk, Ph.D., Robert A. Kruger, M.S., Charles A. Mistretta, Ph.D., Mitchell M. Goodsitt, M.S., Chorn-Gang Shaw, M.S., Stephen J. Riederer, S.M., David C. Flemming, M.D.*, Joseph F. Sackett, M.D., and Andrew B. Crummy, M.D.
Departments of Radiology and *Anesthesiology
University of Wisconsin
Madison, Wisconsin, U.S.A.

ABSTRACT

Two fluoroscopic images which differ by only one variable, e.g., time or x-ray energy, are subtracted in a dedicated real-time digital processor in order to isolate images of injected contrast medium. This technique is applied to dynamic angiography and ventriculography using a peripheral venous injection.

**Work supported by National Science Foundation Grant #APR 76-19076

A real-time video image processor has been developed for the purpose of generating a variety of generalized subtraction images. By subtracting two digitized fluoroscopic images which have only one variable, e.g., time or x-ray energy, one can isolate a subset of the transmission information represented by the terms in a Taylor series of transmission functions[1] (Figure 1). Time-dependent and K-edge energy subtraction are examples of these variables.

The video processor[2]* contains an analog input section, a digitizer, three 256 x 256 x 13 bit memories and dedicated processing hardware circuitry.

X-rays produced by a conventional tube are heavily filtered with materials like cerium, iodine, or brass in concentrations of 100-400 mg/cm^2 depending on the particular application (Figure 2). A plumbicon views the output of the CsI image intensifier and sends a signal to a 256 level automatic gain control circuit (AGC) which standardizes the input to a logarithmic amplifier which preceeds a 10 MHz sample-and-hold and analog-to-digital convertor pair (ADC). The digital processor has three identical data channels to handle the three beam energies we use in our K-edge work[3] (Figure 3). Images are circulated through the memories, processed and displayed in real time. Any linear combination of the three images can be selected.

We have been studying the vessels and the cardiac chambers of the dog following a peripheral venous injection utilizing time dependent subtraction. A mask is integrated over 1/2 second prior to the injection of a 10 cc bolus of Renografin 60 into an extremity vein of the dog. During and following the injection, live images are continually digitized and subtracted from the mask at

*Designed and assembled at the University of Wisconsin Physical Sciences Lab.

CLASSIFICATION OF TRANSMISSION IMAGE COMPONENTS

$I(x,y,\bar{z},\bar{t},\bar{E}) + x,y$ DERIVATIVES	RADIOGRAPHY
$\frac{\delta I}{\delta Z}$	TOMOGRAPHY
$\frac{\delta I}{\delta t}$	TIME DEPENDENT SUBTRACTION
$\frac{\delta I}{\delta E}$	K-EDGE IMAGING
$\frac{\delta^2 I}{\delta Z \delta t}$	TIME DEPENDENT SUBTRACTION TOMOGRAPHY
$\frac{\delta^2 I}{\delta Z \delta E}$	K-EDGE TOMOGRAPHY
$\frac{\delta^2 I}{\delta E \delta t}$	TIME DEPENDENT K-EDGE IMAGING
$\frac{\delta^3 I}{\delta E \delta Z \delta t}$	TIME DEPENDENT K-EDGE TOMOGRAPHY

Figure 1: Classification of transmission image components as derivatives of I(x,y,z,t,E).

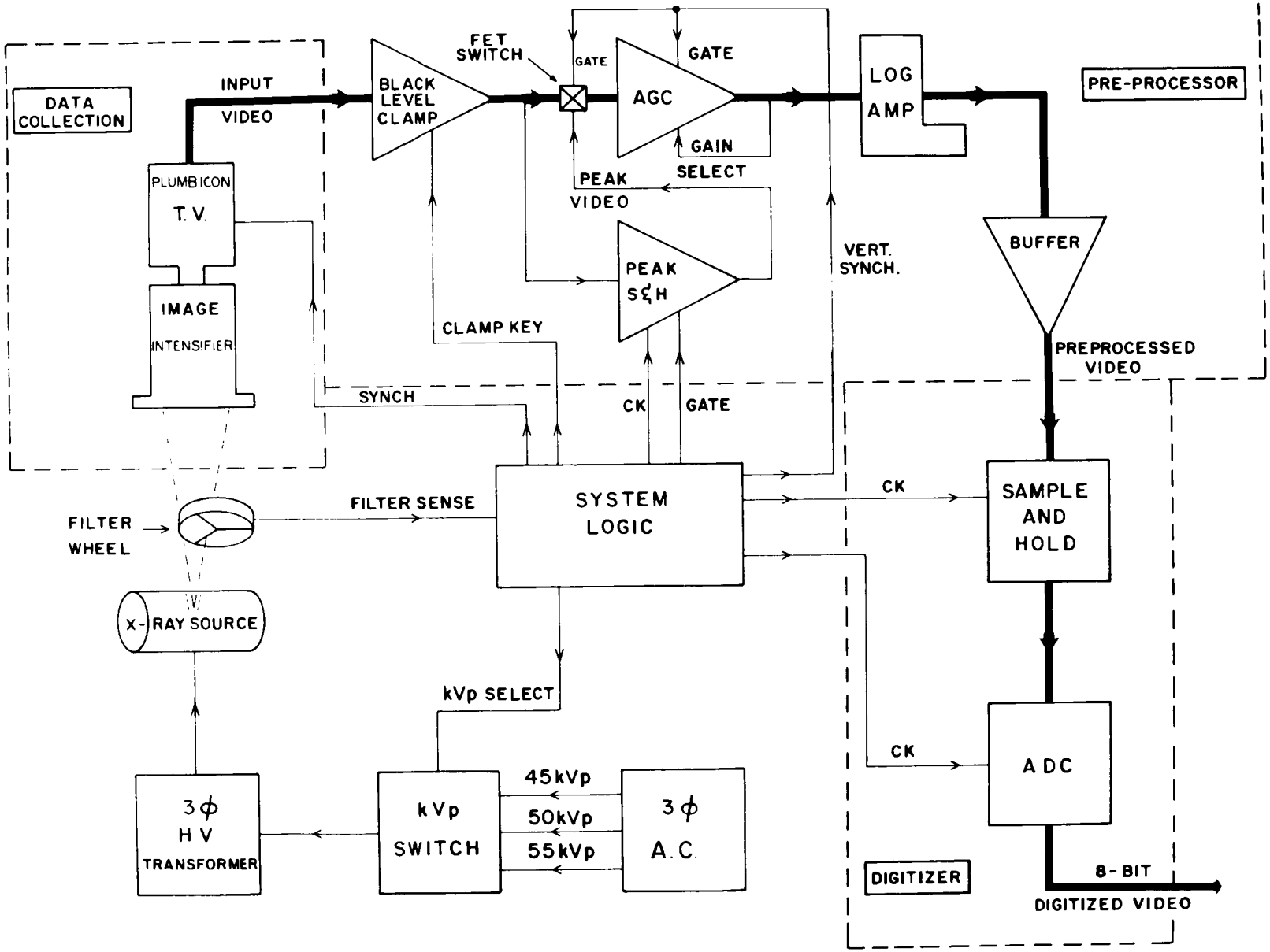


Figure 2: Processor input stage block diagram.

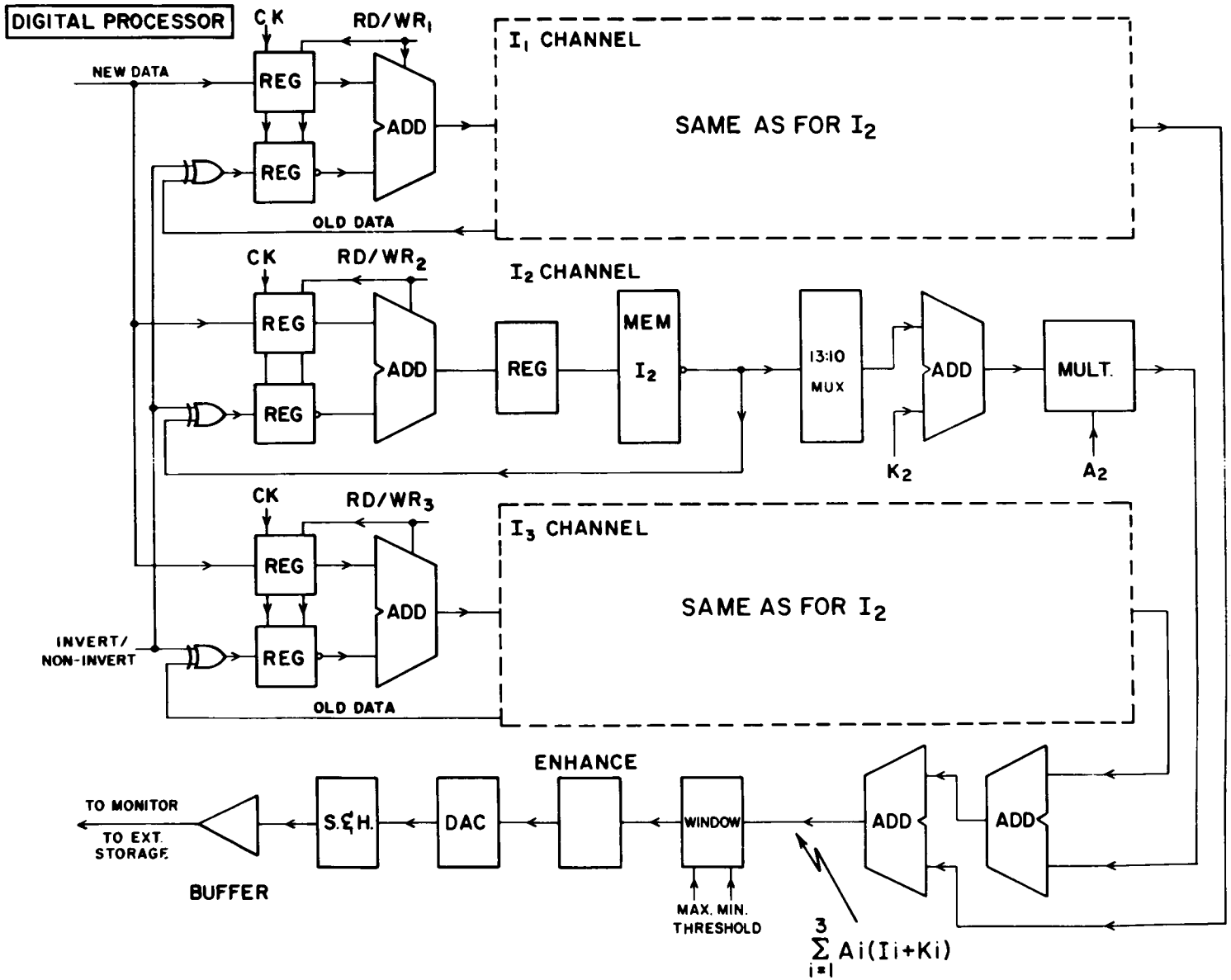


Figure 3: Processor digital section block diagram.

video rates and the results are stored on a magnetic tape. Identical 10 cc injections viewed with conventional fluoroscopy and with time subtraction using logarithmic amplification and 100 mg-cm^{-2} of cerium filtration are illustrated in Figure 4A and 4B. The subtracted images more clearly display the cardiac chambers. A composite of several digital subtraction images made following the intravenous injection of 10 cc's of contrast agent in the foreleg (Figure 5) and the hindleg (Figure 6) demonstrate cardiac dynamics and pulmonary blood flow.

Similarly, peripheral arteries can be studied. When motion is not a problem, 1/2 second exposures are integrated and subtracted from the pre-injection mask to produce a time sequence of arterial and venous images (Figure 7).

With appropriate positioning for imaging, one may follow the progression of a bolus from the site of injection to the heart and then to the periphery (Figure 8).

The 256 x 256 resolution is probably not sufficient for most diagnostic tasks. Therefore, we are installing a 512 x 512 imaging system.

The video tape recordings of blood flow from which the digitized subtractions of Figures 4 through 6 were made are viewed in real time and at 1/4 speed. The videotapes show the turbulent mixing of the opacified and non-opacified blood in the right side of the heart. Opacification of the left-sided chambers is more homogeneous because of the mixing in the right-sided chambers and the pulmonary vessels. The size of the ventricles and the thickness of the myocardium can be determined from these images. Since transmission data is available in digital form, quantitative plots of the ventricular volume versus time can be obtained via videodensitometry without making assumptions about ventricular shape. The dynamics of cardiac contraction are well displayed and the resolution should be sufficient to determine dyskinetic or akinetic areas.

The videotape of the pulmonary vessel opacifications suggests that with the 512 x 512 matrix it will be possible to detect pulmonary emboli utilizing a peripheral venous injection of contrast agent.

Based on our experience with the dog, we estimate that the exposure required for human work will be in the order of 1 to 2 R for an examination of 5 seconds duration. This assumes a 20 cm chest and a 20 cc injection of Renografin 60 into the patient's arm.

The time subtraction technique requires both a pre- and post-injection image. The K-edge energy subtraction technique does not require the pre-injection image. In the K-edge subtraction technique, three quasimonoenergetic beams are prepared by filtration with iodine, cerium, and brass filters which are held in a rotating wheel (Figure 9). The image processor solves for the iodine distribution and displays it 5 milliseconds after the third image is obtained. By varying multiplication constants controlled by up and down counters, bone can be made to pass from black through zero to white as the operator views the image in real time. During this process, the iodine remains white. Very small concentrations of iodine which may not be detectable with routine radiography can be outlined with the K-edge subtraction technique (Figure 10).

We are now doing our K-edge imaging at the rate of 30 Hz. Preliminary results indicate that currently the image quality of dynamic events with the K-edge subtraction is not as good as that with the time subtraction images. However, their freedom from patient motion artifacts is a distinct advantage and warrants further work. Instrumentation is being constructed to allow post exposure synthesis of geometric tomograms[5,6] and K-edge energy subtraction tomograms.[7]

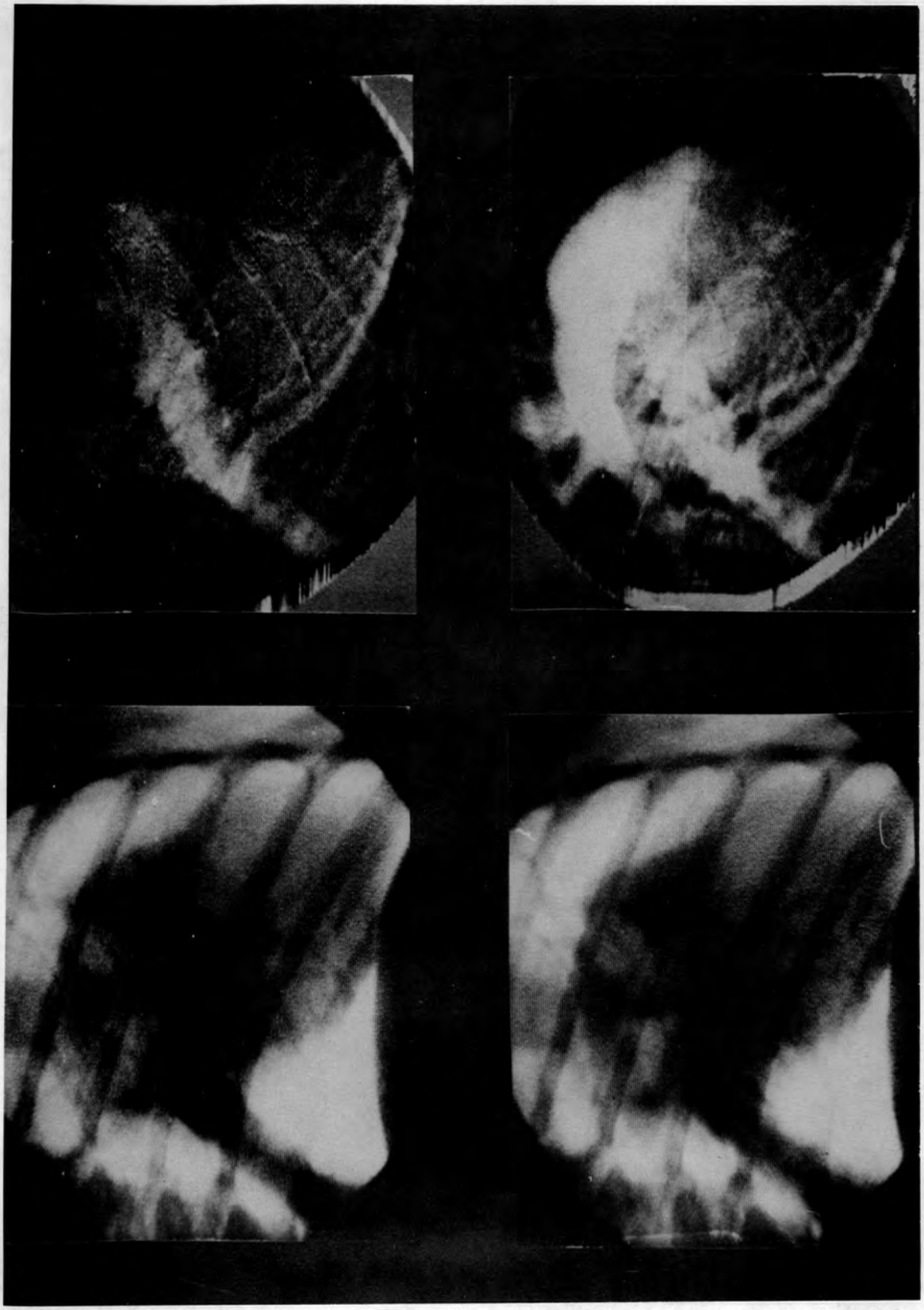


Figure 4A: Television fluoroscopy (left) compared with time-dependent subtraction (right) using intravenous contrast agent in a dog. Upper pair: bolus entering the right heart from the inferior vena cava. Lower pair: the right ventricle and pulmonary artery are opacified and the left ventricle is beginning to fill with contrast.

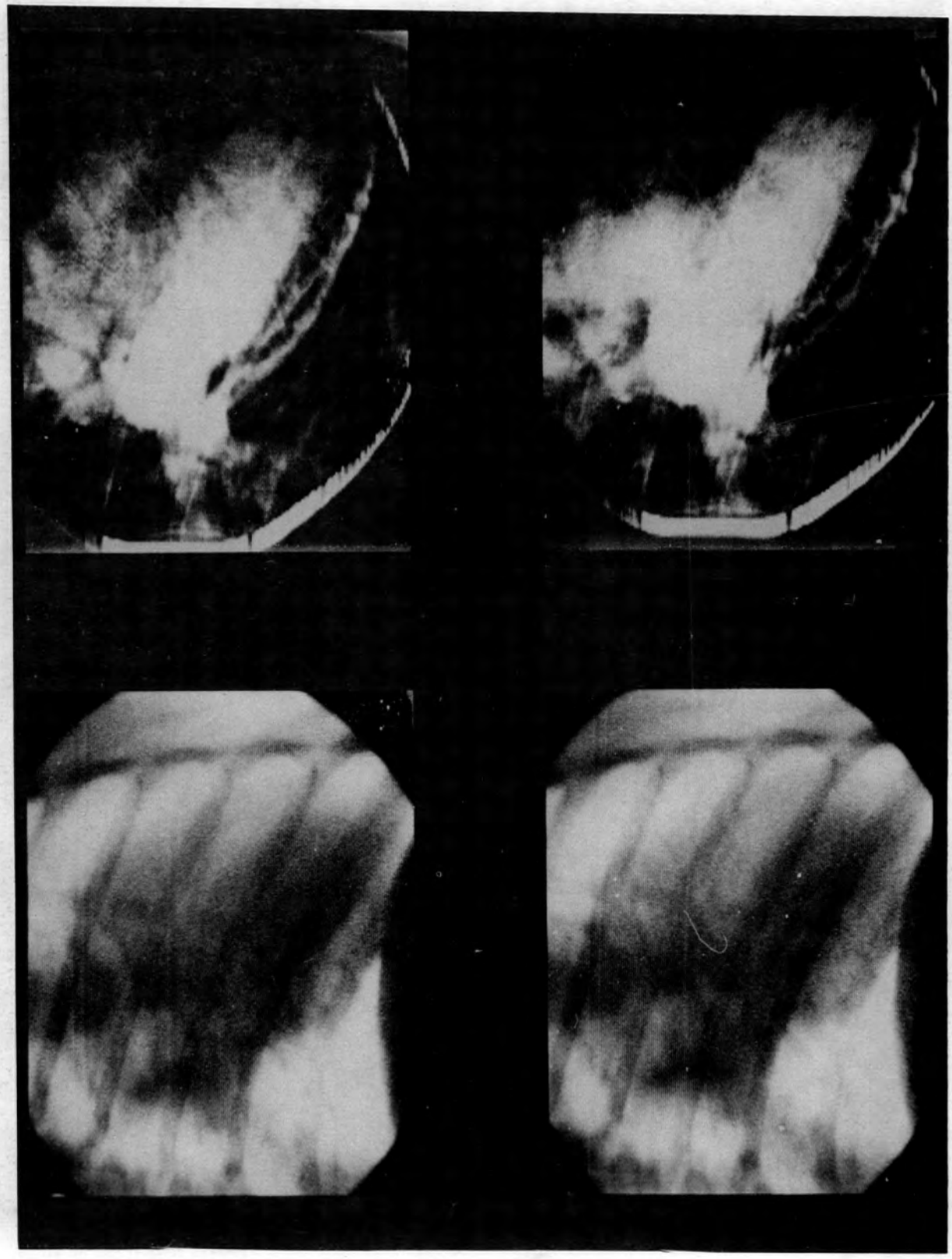


Figure 4B: Later in the cardiac cycle; Upper pair: left ventricular diastole.
Lower pair: left ventricular systole.

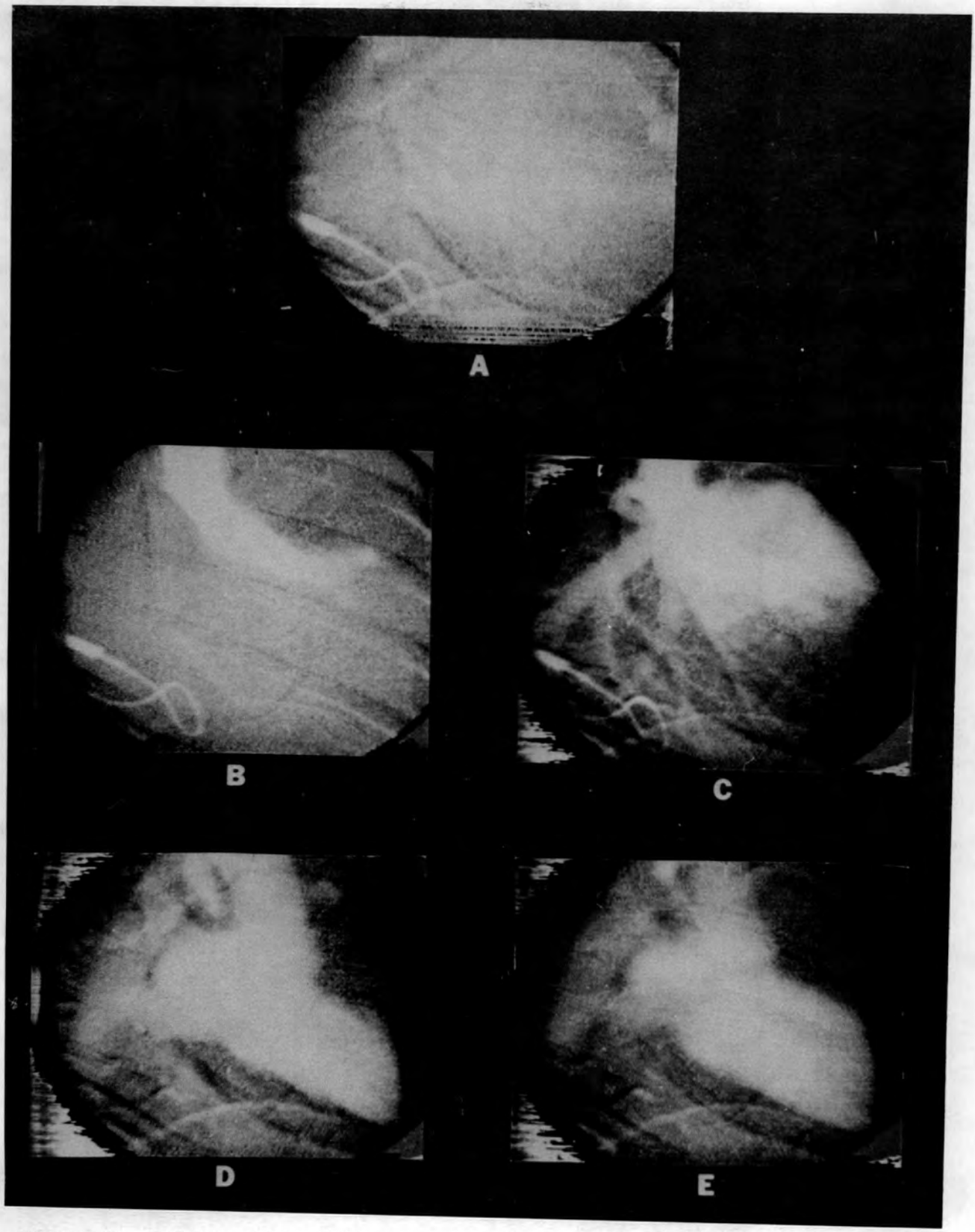


Figure 5: Iodinated blood flow in a dog heart, following bolus injection of contrast in a foreleg vein. A) Bolus entering the vein in the foreleg; B) Contrast agent entering the right ventricle; C) Opacification of the pulmonary arteries; D) Left ventricular systole; E) Left ventricular diastole.

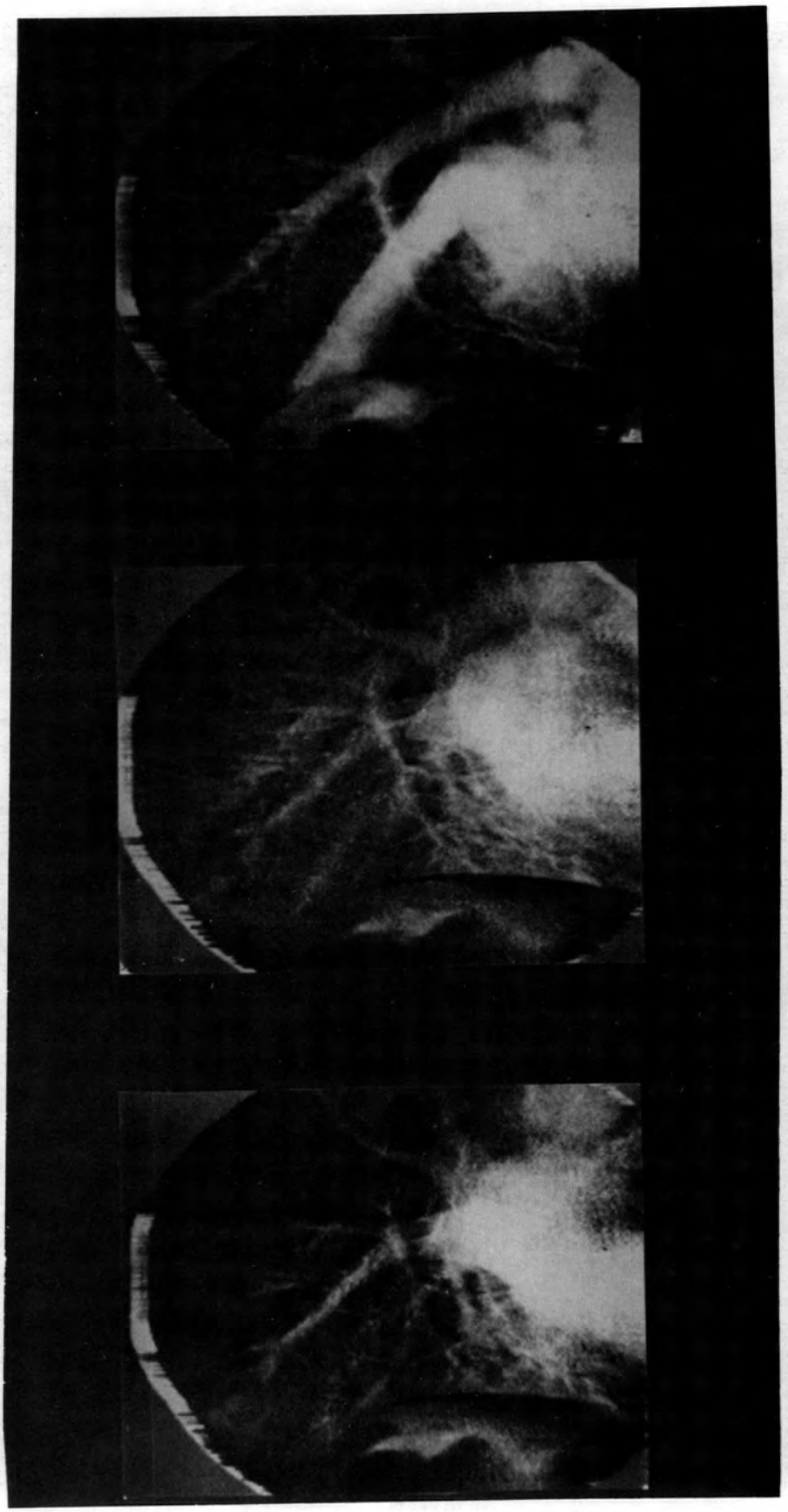


Figure 6: Pulmonary arteriograms in a dog. A) Pulmonary arteries opacifying; B) Pulmonary veins and left heart opacified, emptying of opacified blood from the pulmonary arteries; C) Slightly later in the cycle.

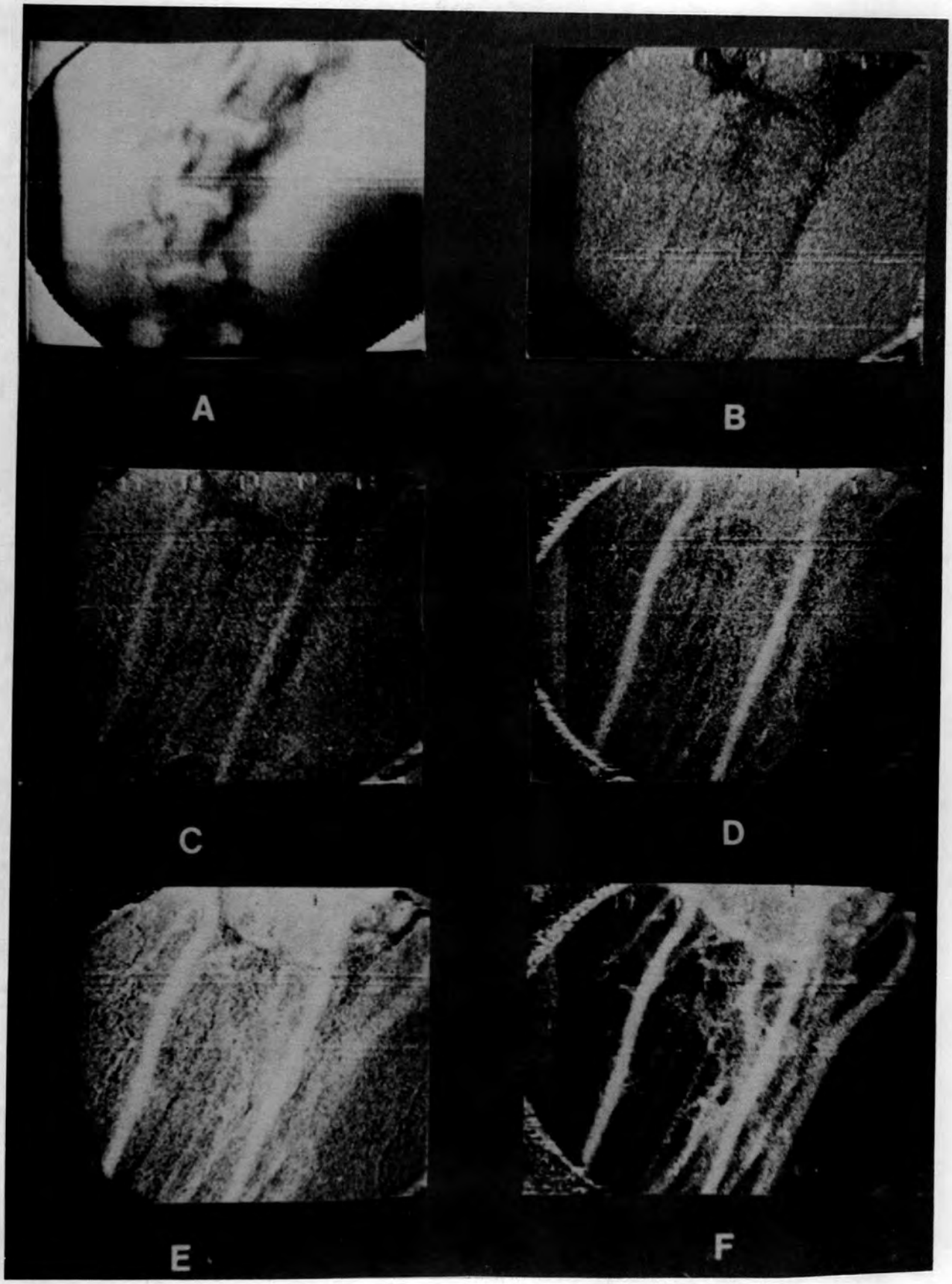


Figure 7: Angiograms of the cervical area of the dog: time sequence of 1/2 s integrations. A) Digital radiographs used as a mask; B) Pre-injection subtraction image; C) Early fill of the carotid arteries; D) Early fill of the vertebral arteries; E) Early venous return; F) Later stage in venous opacification.

Figure 8: Angiographic collage extending from the heart to the skull of a dog.

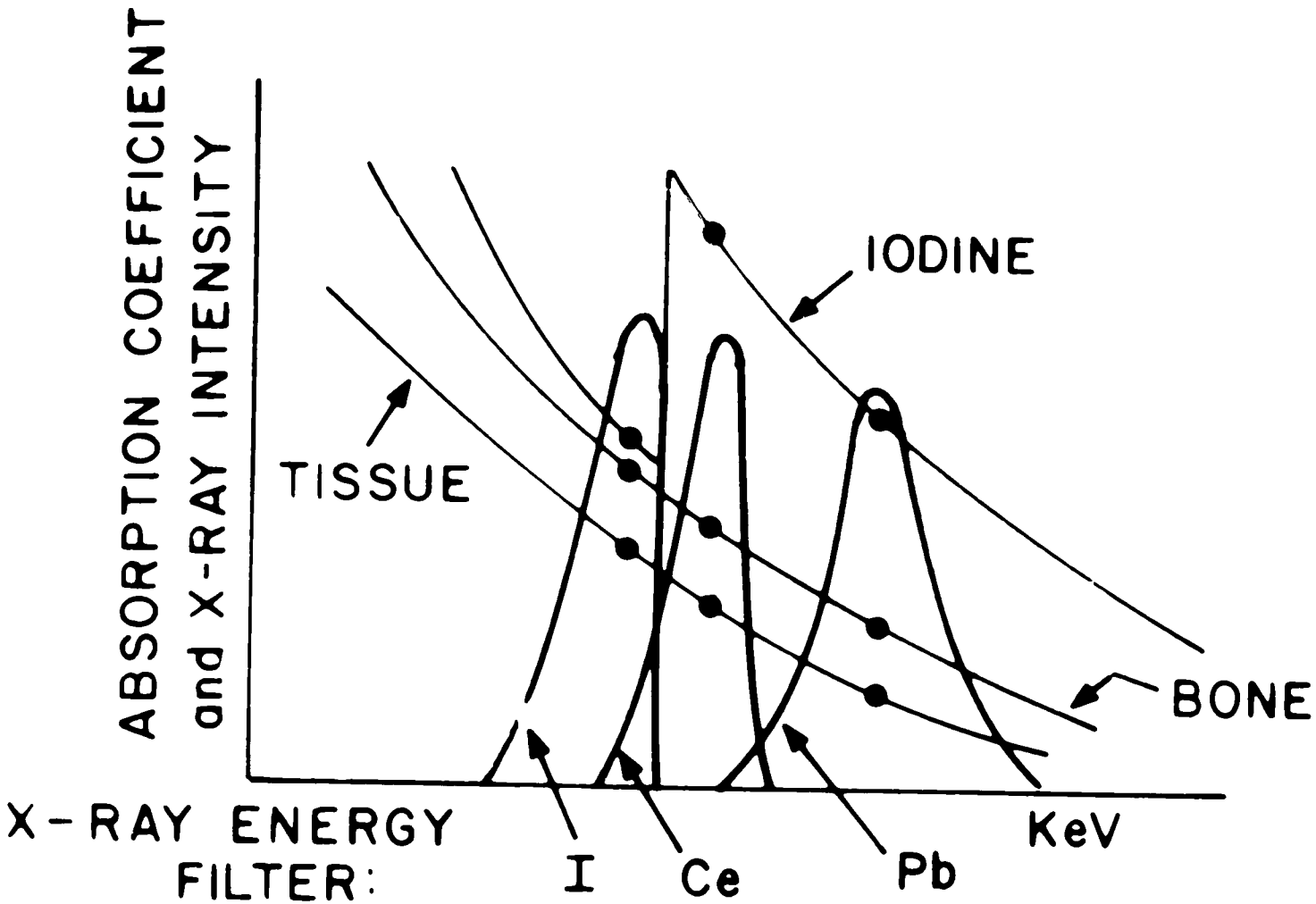


Figure 9: Absorption coefficient and x-ray spectra vs. energy involved in K-edge energy subtraction images.

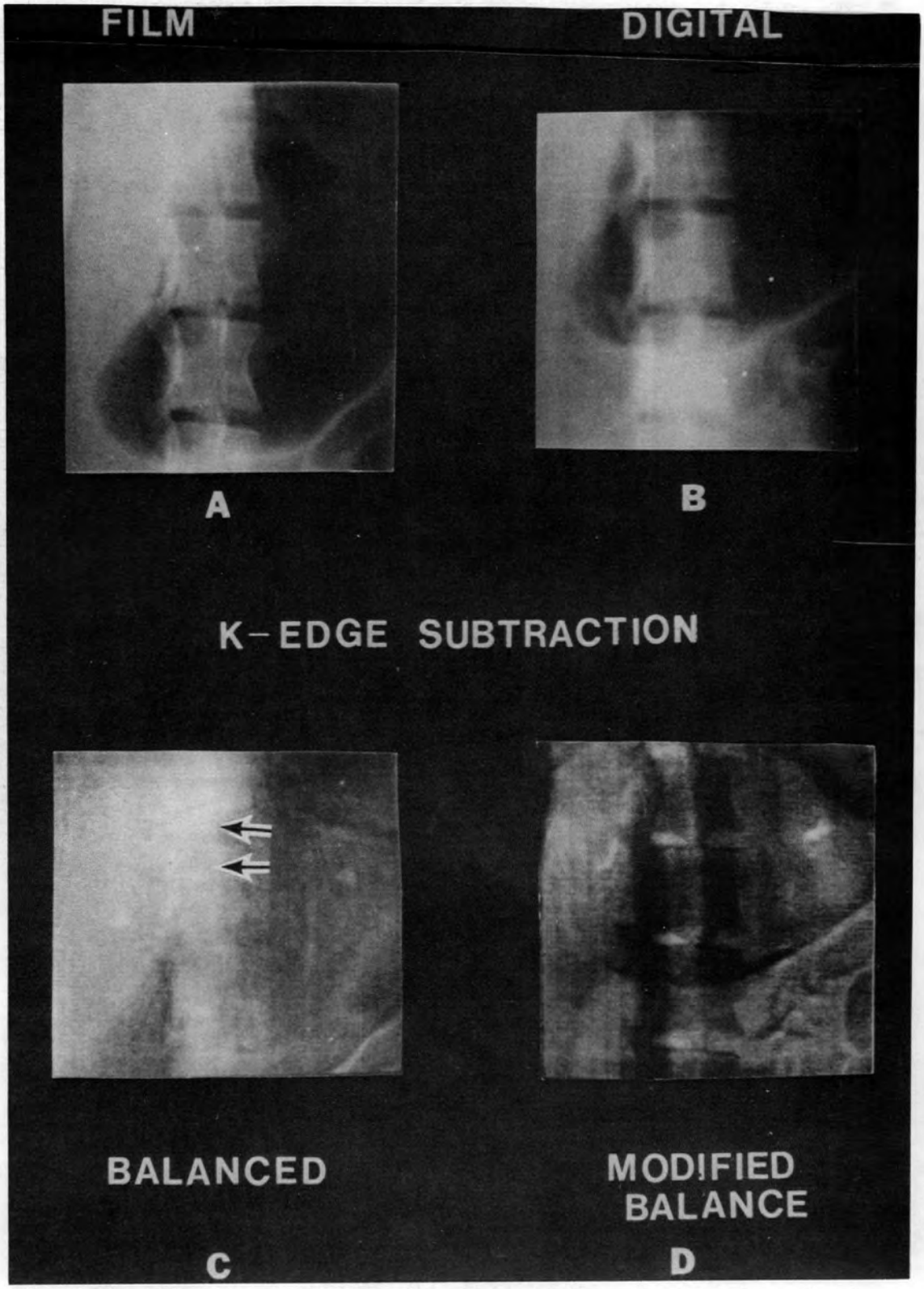


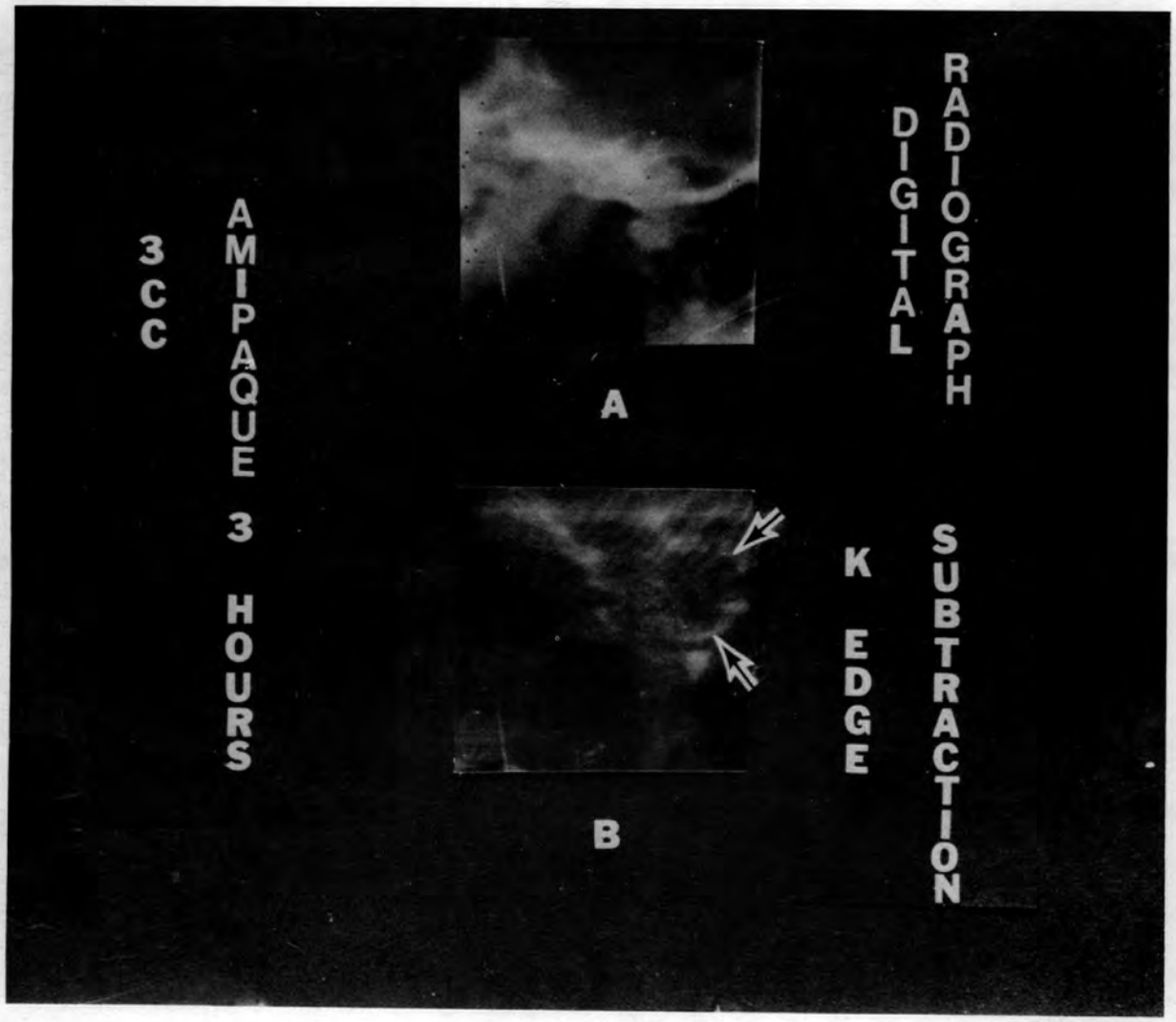
Figure 10: Three cc's of metrizamide (a water soluble, iodinated contrast agent) had been injected into the cistern magna of a rhesus monkey 3 hours previously.

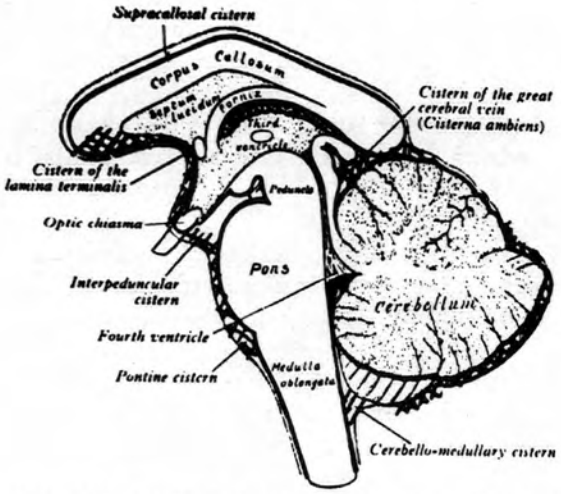
A) A standard radiograph of the lumbar vertebral column of a monkey. No contrast agent can be definitely identified.

B) In the digital picture the spinal cord can be seen similarly by the opacified cerebral spinal fluid.

C) The image has been manipulated to subtract the bone; however, the iodine is not optimally displayed.

D) Further manipulation has resulted in enhancement of the image of the spinal cord.





7.164 A diagram showing the positions of the principal subarachnoid cisterns. *Red*: pia mater. *Blue*: arachnoid mater.

C

Figure 11: A) Fluoroscopic image of iodine (metrizamide) in a monkey skull 3 hours post-injection.
B) K-edge image of A shows the cerebellum outlined by the contrast agent.
C) Sketch of human cisternal anatomy showing the iodine (cross hatched) as it appears in the K-edge image.

REFERENCES

1. C. A. Mistretta, "The Use of a General Description of the Radiological Transmission Image for Categorizing Image Enhancement Procedures", Optical Engineering, 13, #2, 134 (March/April, 1974).

2. R. A. Kruger, C. A. Mistretta, J. Lancaster, T. L. Houk, M. M. Goodsitt, C. G. Shaw, S. J. Riederer, J. Hicks, J. F. Sackett, A. B. Crummy and D. Flemming, "A Digital Video Image Processor for Real-Time Subtraction Imaging", submitted for publication to Medical Physics.

3. F. Kelcz and C. A. Mistretta, "Absorption Edge Fluoroscopy Using a 3-Spectrum Technique", Medical Physics 3, 159 (1975).

4. R. A. Kruger, M. M. Goodsitt, C. A. Mistretta, A. B. Crummy, J. S. Sackett, S. J. Riederer, T. L. Houk, C. G. Shaw, "Digital K-Edge Subtraction Radiography", accepted for publication in Radiology.

5. N. A. Baily, R. L. Crepeau, and E. C. Lasser, "Fluoroscopic Tomography", Investigative Radiology 9, #2, 94 (March/April, 1974).

6. E. E. E. Hoeffer, "Electronic Synthesis of Tomograms", Medicamundi, 19, #2, 866 (1974).

7. R. A. Kruger, S. J. Riederer, and C. A. Mistretta, "Comments on the Relative Properties of Tomography, K-Edge Imaging and K-Edge Tomography", Medical Physics, 4, #3 (May/June, 1973).

DISCUSSION:

Pullan asked what the concentration of iodine was that was used for the studies. Dr. Houk stated that he did not know the concentration but noted that the procedure used 10 cc's of Renorafin-60 flushed with 10 cc's of saline.

Pullan stated that he had made calculations which indicated that one needs at least 4 mg/ml of iodine before the contrast medium would become visible.

Houk stated that on static views he had successfully visualized 2 mg/cm^2. He also noted that in static studies, they routinely integrated over 8 frames which corresponded to a study time of about 1/2 second.

Price asked about motion artifacts in the K-edge images. Houk noted that there would be motion artifacts generated if there were any movement between the preinjection mask and the current image. He also noted that at the present time no mechanism for gating the system had been considered.

IMPACT OF LARGE SCALE INTEGRATION (LSI) ON MEDICAL RADIONUCLIDE IMAGING

Robert O. Smith
University of Mississippi Medical Center
Jackson, Mississippi, U.S.A.

ABSTRACT

Some dynamic organ radionuclide distribution studies are being limited by the frequency response of the electronic circuits in the detection equipment and display devices. Generally, the smaller such circuits can be made, the greater their frequency capability becomes. Large scale integration of digital and analog circuits is already increasing their speed and will ultimately decrease the prorated functional cost.

Medical radionuclide images are being processed faster and at lower cost by large scale integrated (LSI) digital and analog circuits. LSI may well provide a new impetus to radiology and nuclear medicine similar to that produced on computers in the 1960's by integration of transistor, transistor logic which is commonly referred to as TTL or T^2L. With the increasing gamma detection rate of the large single crystal or the multicrystal devices, the advent of these LSI circuits is none too soon.

During the past three years, the cost of an analog-to-digital (A/D) converter for digitizing the position pulses from one channel of the gamma camera has dropped from $430 (Analog Devices ADC 1103 hybrid module, 8 binary bits in 1.0 μs) to $55 for a dual-in-line package (Burr-Brown ADC 82AG, 8 binary bits in 2.8 μs) or approximately $50 (Precision Monolithics DAC-08, etc., 8 binary bits in 1.0 μs) where 13 chips are mounted on one printed circuit (PC) board. A "do it yourselfer" can now construct an 8 binary bit converter for about $35 even when buying the parts in single quantities. Figure 1 shows the PC board for this converter with the components in place. Both dual-in-line components have 16 pins and their size is indicative of the whole assembly. Figure 2 is the complete schematic and demonstrates how really uncomplicated the converter is. The primary components are the fast digital-to-analog converter, the 8-bit successive approximation register and the comparator. This unit can convert to 6 bits in 4.7 μs, to 7 bits in 5.3 μs and to 8 bits in 6 μs when a 1.5 megahertz (MHz) clock is used. Figure 3 lists the components. The digital-to-analog converter and comparator are the first two items and are manufactured by Precision Monolithics. The next item is the successive approximation register which can be obtained from Advanced Micro Devices, National Semiconductor or Signetics. The remaining components can usually be obtained locally.

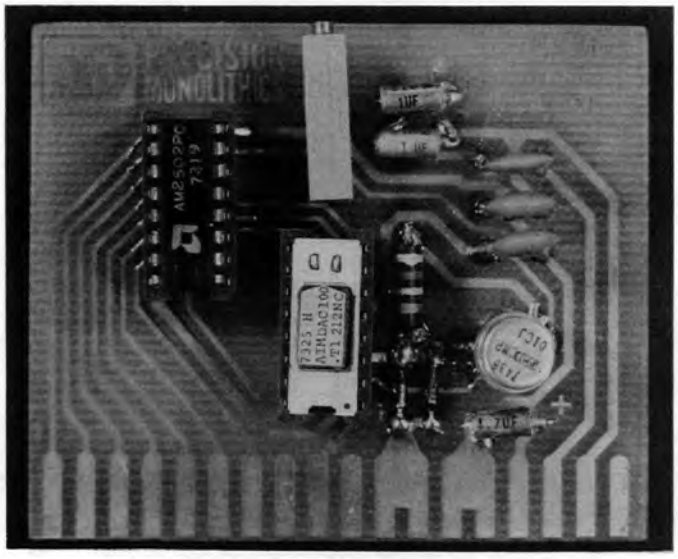


Fig. 1. Photograph of 8-bit Analog-to Digital (A/D) Converter with only 13 components mounted on printed circuit (PC) board of only 6 x 7.3 cm.

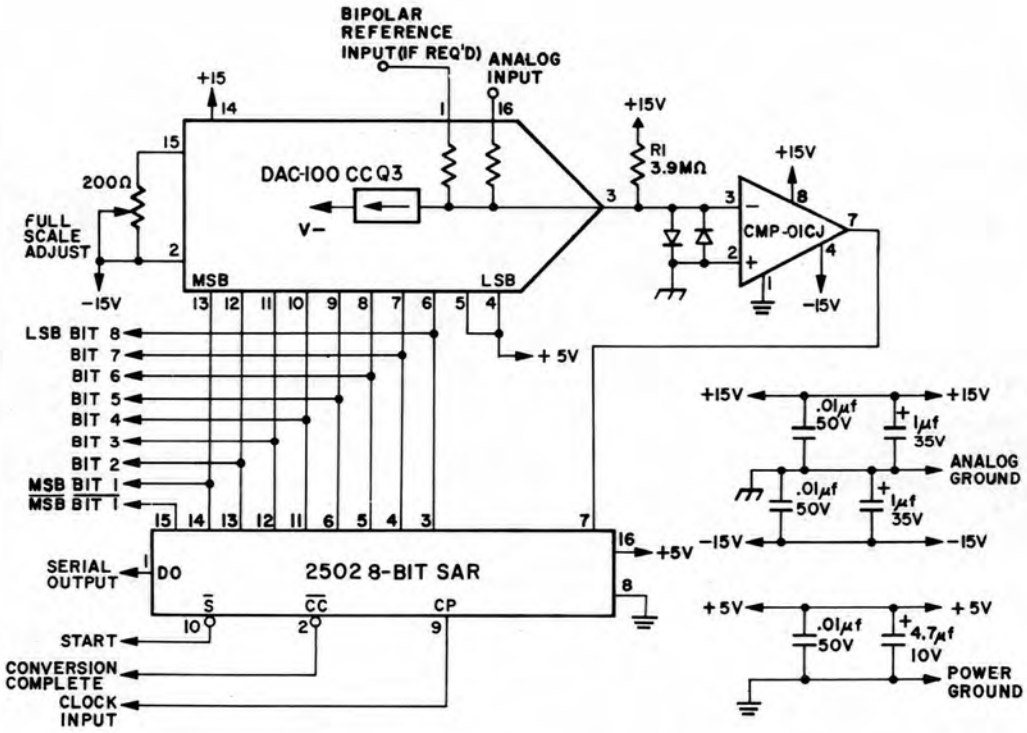


Fig. 2. Complete 8-bit A/D schematic of converter shown in Fig. 1. (From Precision Monolithics, Inc. Application Note #AN-11.)

±0.3% maximum nonlinearity, FS tempco 120ppm/°C

1	DAC-100DDQ3 (or Q4)
1	CMP-01CJ
1	AM2502PC (Advanced Micro Devices) or Equivalent
1	Pot-200Ω Bourns #3006P-1-201
1	4.7 μf CAP- Mallory #TDC475M010EL
2	1.0 μf CAP- Mallory #TDC105M035EL
2	Diode, 1N4148
3	.01 μf CAP-Centralab #CK-103
1	PC Board
1	Resistor 3.9MΩ 5% 1/4W

For ±0.2% maximum nonlinearity, FS tempco 60ppm/°C use DAC-100CCQ3 (or Q4)

Fig. 3. Parts List for 8-bit A/D converter. See text for sources.

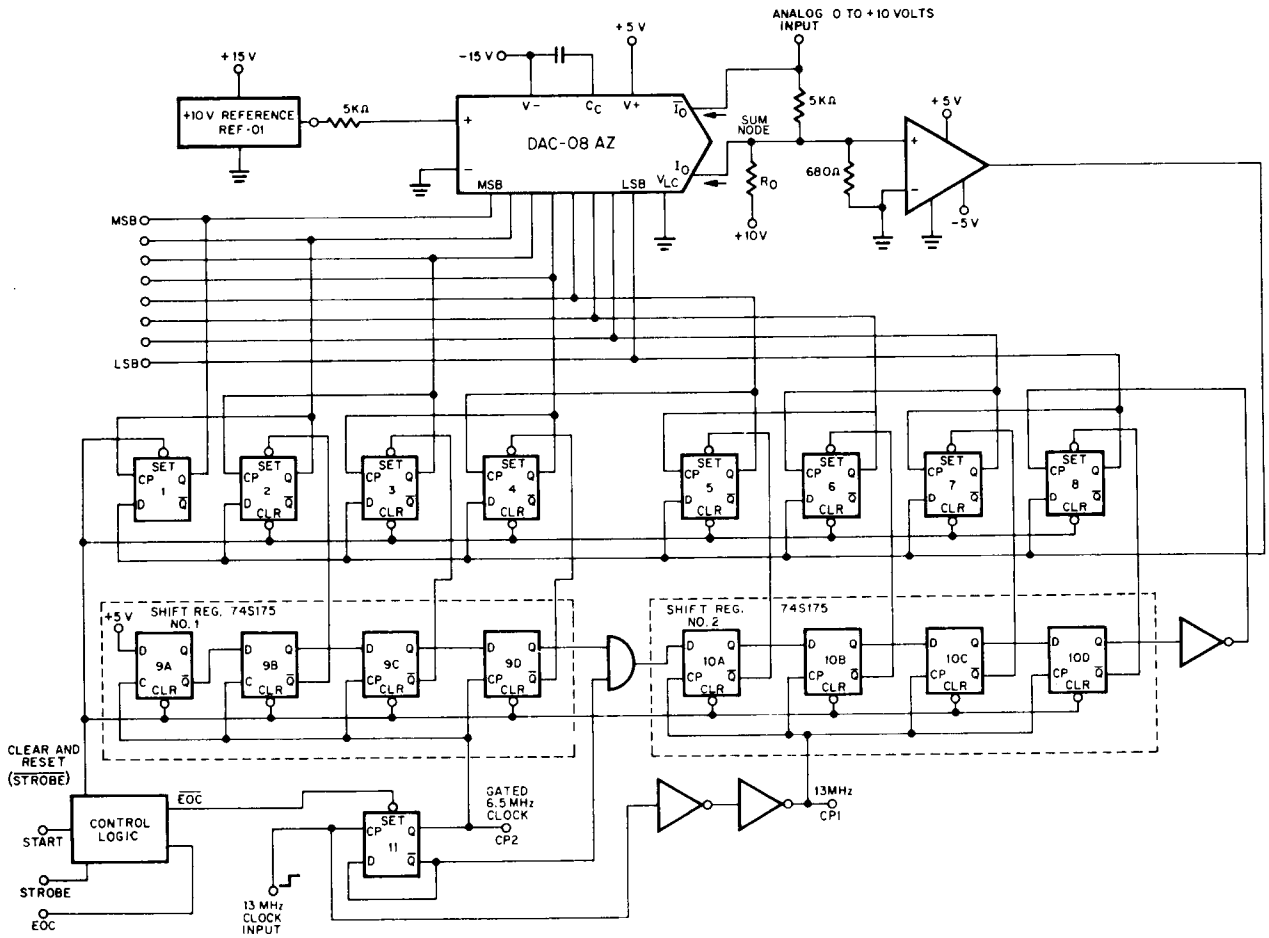


Fig. 4. Simplified schematic of one μsec A/D converter. (From Precision Monolithics, Inc. Application Note #AN-16.)

Figure 4 is a simplified schematic of a faster analog-to-digital converter that will convert to 8 binary bits in one microsecond using a 13 MHz clock. Figure 5 is the complete schematic for the one microsecond converter. There are six Dual D flip-flops, two Quad D flip-flops, one Quad 2-Input Nand gate, one Triple 3-Input Nand gate, one 8-Input Nand gate, the digital-to-analog converter and the comparator for a total of thirteen integrated circuits. Precision Monolithics can supply the digital-to-analog converter and Advanced Micro Devices the comparator. The other logic listed was Schottky because it has the necessary speed and its use also increased the total number of components when compared to the previously described less complicated converter.

With the schematic and parts lists for an A/D converter, construction should not be too difficult if a few rules are followed. PC board layout has always been important with analog-to-digital converters and for high speed designs it is even more important. Several of the more important rules of layout are: First, digital ground must be separated from analog ground. They must have only one common point. Second, digital traces or conductors should not cross or be routed near sensitive analog areas. This is especially important near the sum node. Third, with Schottky TTL logic, the digital ground and the five volts supply conductors or traces should be large and contain provisions for generous bypassing with capacitors. Fourth, the conductor (sum node) from the digital-to-analog converter output to the comparator input should be short. It should also be guarded by analog ground. Fifth, the analog input conductors should be short. Therefore all analog components should be located as close as possible to the edge connector. Lastly, the comparator's outputs should be routed away from its inputs to minimize capacitive coupling and possible oscillation.

Figure 6 shows the typical system connections for such a converter. As shown, end of conversion or EOC is connected to start for continuous conversions. This would not be correct for digitizing camera position pulses. As an option, Strobe is used to clock the parallel answer into an output register at the end of each conversion. The dotted line jumper is the one point at which the digital and analog grounds meet. The four power supply voltages are a positive ten volt reference, negative fifteen volts for power, negative five volts reference and positive five volts for powering the logic.

The detected and digitized scintillation events must be stored for additional processing and display of the results. Storage at this stage appears not to be a problem for most memories used with minicomputers or microprocessors will suffice. When camera devices can approach handling a half million counts per second, then additional computational and storage speed will be required in the acquisition system. Insufficient speed is already a problem with intensity modulated cathode-ray-tube (CRT) displays. For the American Standard Display using 256 points per line and 240 lines per field, the point rate is 5.04 MHz. If the control processor has a 16-bit word length and each dot is either on or off as for alpha-numerics, then the processor can output words to the CRT interface at about one-sixteenth the 5.04 MHz rate. This is because each word can store the on/off condition of 16 dots or locations. When intensity modulation is needed for displaying computed scan information, complications arise in making the data available at the required intervals. For 32 gray levels (5 binary bits) each 16-bit word could supply the intensity for three dots or locations. The processor would have to output data at one-third the 5.04 MHz rate and this is beyond the capability of most micro or central processors. However, if the points per line were reduced to 128 and there were 120 lines per field, the point rate would reduce to 1.26 MHZ and the word rate would be only one-third

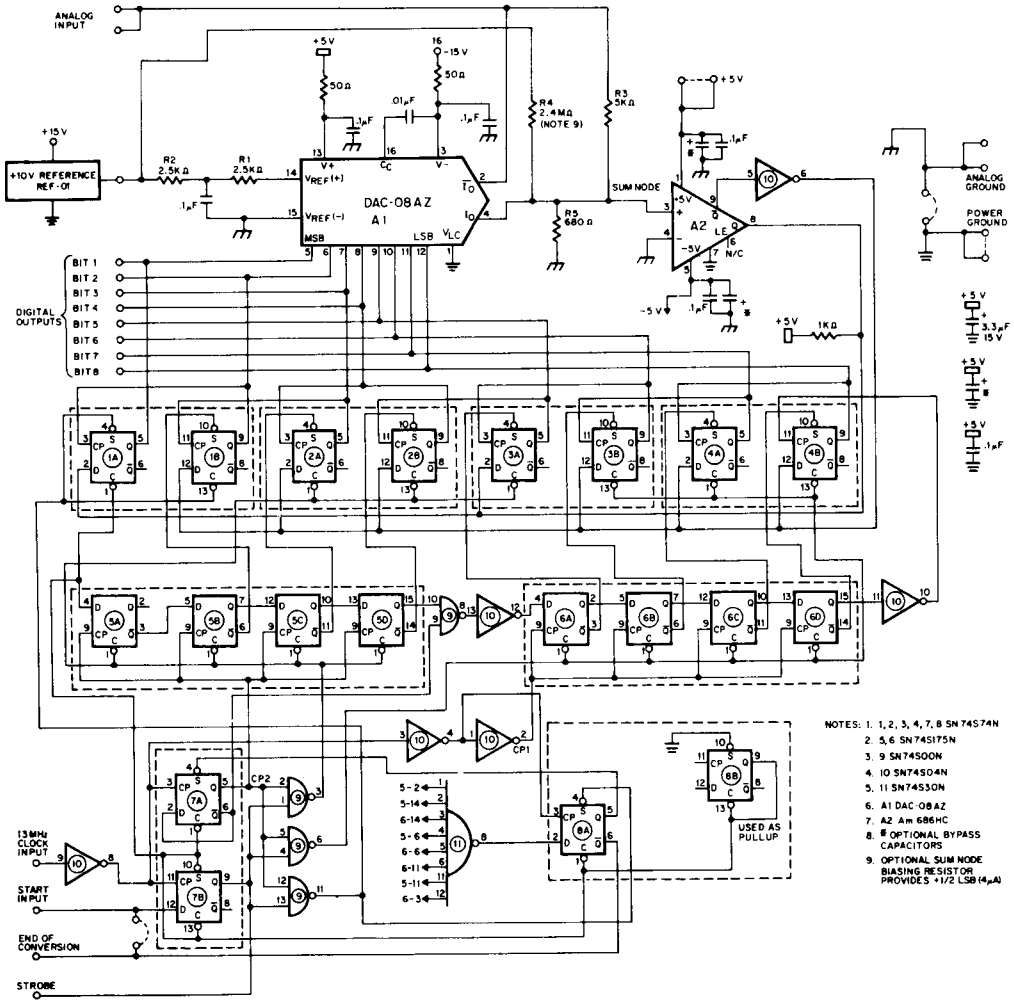


Fig. 5. Complete construction schematic of converter shown in Fig. 4. Additional details can be found in Application Note #AN-16.

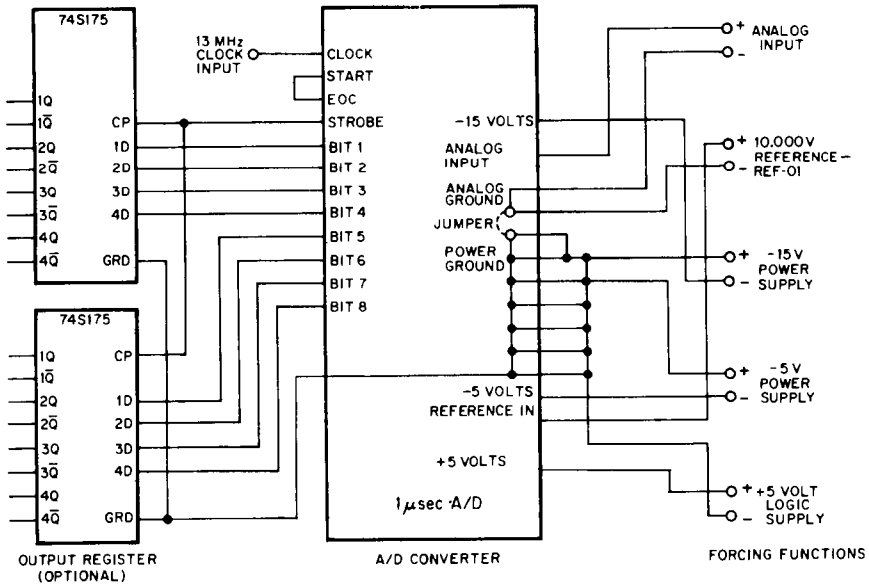


Fig. 6. Typical system connection of converter shown in Fig. 5. The two 74S175 registers are optional buffers to allow conversion to resume while transferring the previous data. Also from Application Note #AN-16.

of 1.26 MHz and this is feasible with most of the later processors. However, when the goal is 32 levels of intensity for each of the three primary video colors, complications as to speed do arise. This requires that data for each point be output from storage in less than 200 nanoseconds. The refresh type (dynamic) of random access memory is not currently this fast though some versions are approaching this range. Though bipolar memory appears adequate, it is in limited supply at high prices. For the moment and considering the cost, intensity modulated color CRT displays may have to use other alternatives such as a 128 by 128 array, interpolate on the fly or perhaps go to other storage devices.

One such a device is shown in Figure 7. This is a charge coupled device (CCD) manufactured by Fairchild which has a 9216-bit dynamic shift register memory consisting of 9 registers of 1024 bits each and is intended for 8-bit byte plus parity operation. Operating frequency is 100k to 2 MHz and price is about $31.00 each in quantities of 25 or more. An even larger shift register memory is shown in Figure 8. At the maximum operating frequency of 5 MHz, it provides a data rate of 20 megabits because data is available 4 bits parallel. This parallel output provides some unique possibilities for on the fly interpolation to smaller display cells. This is a 22 pin dual-in-line plastic device with 128 recirculating registers of 128 bits each. The 128 registers are divided into 4 groups of 32 each for parallel addressing, input and output. It is listed as a 16,384-bit dynamic line addressable random access memory or LARAM. Both this memory and the previous one are TTL compatible with 3 state outputs but there are special requirements. The clock must be two phase with zero to positive 12 volts drive capability. Whenever the stored time of any bit in any register has exceeded the specified maximum refresh period, that register must be purged before it can be used to store data. Purging is accomplished by clocking the register a minimum of 512 clock cycles. When the chip is first powered up, the entire 16k memory must be purged by addressing each line and clocking each line a minimum of 512 cycles. However, the refresh period at reasonable temperatures appears long enough to allow stopping the clock at the end of each horizontal CRT scan line and also perhaps during vertical retrace. There is extensive research in CCD's at the present time and the future may show them to be very useful devices.

In spite of the rather slow speed of the early microprocessors such as the Intel 8080 and the Motorola 6800, they have been useful in many ways including training for more advanced devices. When gamma camera data have been processed, minicomputers have been the preferred devices for counting the detected gamma events. While in the data acquisition mode, processing of the data cannot be accomplished because too much time and consequently counts would be lost in switching back and forth between the acquisition and processing programs. The decreasing prices of microprocessors and their memory makes them attractive as dedicated data acquisition devices. The least expensive and easiest to implement microprocessors are those of eight binary bits but this word size limits them to 255 counts/pixel which in turn limits the total counts to $<$200k in a 32 x 32 image. Although 16-bit microprocessors are available, a more practical solution may prove to be use of three 4-bit microprocessor slices for an even faster processor combined with a 12-bit word length. The acquired data could be stored on perforated paper tape, magnetic tape or floppy disc for later rapid entry into the minicomputer which could then be used full time for scan processing and administrative functions. Manufacturers of image display devices were among the first to utilize microprocessors in their equipment but primarily for flipping the arrays for correct orientation on the CRT screen. Only the emitter coupled microprocessors appear to have adequate speed for controlling data flow to a high resolution (256 x 256) composite color video display where each

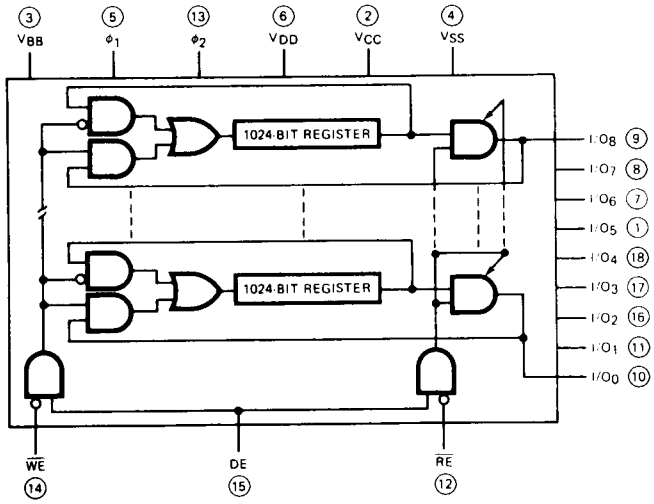


Fig. 7. Block diagram of CCD450 9216-bit dynamic shift register memory consisting of nine registers of 1024-bits each. (From MOS/CCD Data Book by Fairchild Semiconductor.)

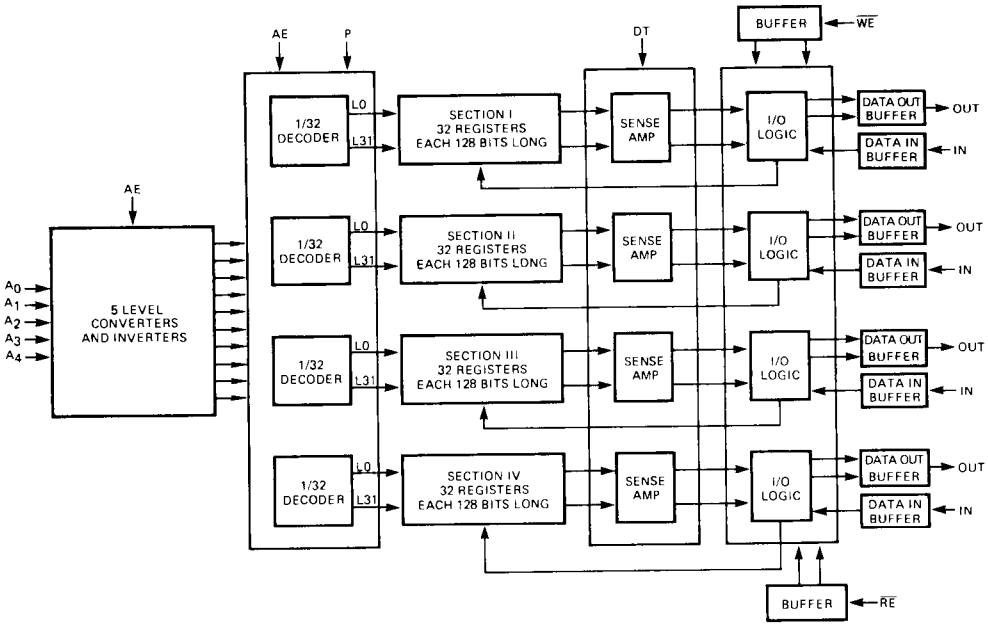


Fig. 8. Block diagram of CCD460 16,384-bit dynamic line addressable random access memory (LARAM). (From MOS/CCD Data Book by Fairchild Semiconductor.)

primary color could have 32 levels of intensity. The speed of emitter coupled logic presents problems in laying out the circuitry to achieve pulse synchronization and to minimize crosstalk. Here again the 4-bit microprocessor slices may become sufficiently fast and yet not present the layout problems. When the data processing is relatively simple, i.e. 9-point averaging or summing of scans, the read only memory (ROM) of the display microprocessor might even be programmed to handle these tasks prior to displaying the data and thus eliminate the need for a minicomputer.

Solid state memory has largely supplanted magnetic core memory in display and processing applications. Although generally faster than core, speed is a continuing problem with solid state memory when large arrays are required to improve image resolution by decreasing pixel size. Manufacturers are currently decreasing the access time of random access memories (RAM) of both the static and dynamic types so that cycle times of less than 200 nanoseconds may be common by this fall or winter and without using the expensive bipolar configuration. This time frame should also complete the development of families of 4-bit slices.

Only a minimal impact is presently seen from LSI devices on nuclear medicine. The first new equipment with these circuits was actually higher in price because of its extended capability and the need to recover research and development costs. As quantities that are used increase, prices should decrease which should interest newcomers in processing nuclear medicine information. They will not be bound by the format of old equipment and could well eclipse those now in the forefront who may not readily adapt to change. However, for those with previous experience, who like to design and construct their own equipment, LSI should be a boon because more logic functions can be installed on a single PC board than was previously possible.

ACKNOWLEDGEMENT

Some of the figures are from application notes supplied by Precision Monolithics, Inc., 1500 Space Park Drive, Santa Clara, California 95050. The data on the charge coupled devices are preliminary and were supplied by Fairchild Semiconductor, 464 Ellis Street, Mountain View, California 94042.

A portion of this research was supported by National Institutes of Health Research Grant HL-19289.

DISCUSSION:

Winspur asked about the contents of 512 x 512 memory discussed in the paper. Smith noted that the design specified 5 bits resolution for each of the three colors.

Winspur noted that the Grinnell display system provided 512 x 512 with 10 bits of brightness or color levels.

MULTIPLE CATEGORIES FROM MEDICAL IMAGES: A SUPPLEMENT TO ROC ANALYSIS

Alan B. Ashare, M.D.
Wright State University School of Medicine
Dayton, Ohio USA

ABSTRACT

ROC (receiver operating characteristic) analysis has provided a technique to compare the results of one imagining technique to another.[1] However, under some circumstances, two imaging techniques may appear, by ROC analysis, to be of equal value. However, if the abnormal category is further delineated into specific diagnoses, then a difference may be seen in the two imaging techniques. This multiple category approach yields an information content metric which compares the diagnostic capabilities of different imaging techniques.

INTRODUCTION

Medical images are obtained to help the physician in the diagnostic process. There have been many improvements in these imaging techniques in the past decade, and methods are needed to evaluate these improvements. One method is ROC analysis[1] which has proven to be of great value. ROC analysis considers a binary decision, usually normal versus abnormal. Abnormal medical images in an applied clinical situation often suggest one of several specific diagnoses, however, and ROC curves describing the detectability of abnormality do not reveal the usefulness of an imaging technique in providing differential diagnoses. This problem has been referred to as the difference between detectability (the ability of a system to detect the presence or absence of a signal) and recognizability (the ability of a system to locate, recognize, or identify a signal).[2,3,4]

Expanding the Decision Matrix

The traditional decision matrix (2 x 2) can be expanded into an N x N matrix where N is the number of different diagnoses possible. To illustrate this expansion, thyroid scintigrams were considered. The possible diagnoses included: normal, single cold nodule, single hot nodule, multinodular gland and heterogeneous gland. The actual disease state is located on the ordinate of the matrix and the observed disease state is located along the abscissa. The entries to the matrix are completed in the same fashion as the 2 x 2 decision matrix.

The following are the results from an experiment done with thyroid scintigrams by four physicians.[5]

Table 1. Technique I

		OBSERVED				
		Normal	Single Cold	Single Hot	Multi	Hetero
Actual	Normal	115	8	1	1	3
	Single Cold	2	46	0	3	5
	Single Hot	0	2	17	3	2
	Multi	0	5	0	23	4
	Hetero	2	1	0	2	23

Several things can be noted from this matrix, sometimes called a stimulus-response matrix or a confusion matrix. By adding the entries in each row and then dividing by the total number of entries in all rows, the a priori probability for each diagnosis can be calulated. Again this is similar to the 2 x 2 matrix where the false negative studies plus the true positives divided by the total number of entries is equal to the probability that a signal is present (p(s)).

Accuracy

Accuracy in the 2 x 2 matrix is equal to the sum of the true negative studies and the true positive studies (the sum along the major negative diagonal) divided by the total number of entries; a similarity can be noted in the 5 x 5 matrix where accuracy is equal to the sum along the major negative diagonal divided by the total number of entries.

Accuracy = (115 + 46 + 17 + 23 + 23)/268 = 89%

Looking at the columns, it can be noted that only four abnormal scintigrams were read as normal and that these were equally split between the diagnoses of single cold nodule and heterogeneous gland. From this set of data we can see that if a scintigram interpretation is normal, there is a 96% probability that this is correct, 2% probability that it is either a single cold nodule or a heterogeneous gland. Furthermore, there is no chance that the diagnosis is either a single hot nodule or a multinodular gland. A similar analysis can show the usefulness of the diagnosis of a single hot nodule. Only one out of eighteen scintigrams were read incorrectly; therefore, the other three diagnoses can be ruled out (since the probability is 0%).

In the case of the diagnosis of heterogeneous gland, there is a 62% chance that it is correct, an 8% chance it is normal, a 14% chance it is a single cold nodule, a 5% chance it is a single hot nodule, and an 11% chance it is a multinodular gland.

Another technique is used to image the same patients and the following matrix is obtained.[4]

Table 2. Technique II

	OBSERVED				
	Normal	Single Cold	Single Hot	Multi	Hetero
Normal	110	4	0	4	10
Single Cold	15	15	6	7	13
Single Hot	0	1	18	1	4
Multi	14	7	0	1	10
Hetero	14	2	1	3	8

Comparison of the Two Techniques

How are these two techniques to be compared? One possibility is to compare the accuracy measures.

Accuracy

Technique I = 89%

Technique II = 78%

The problem with accuracy measures has already been mentioned by other investigators.[6] Basically, the problem is that this measure is biased when the a priori probabilities of each diagnosis are very different. However, I believe that a more unbiased measure can be derived from classical information theory.

Information Theory

The patient is considered to be the sender of a signal, the physician is the receiver, and the medical image is the channel for transmitting the signal. The information transmitted [T(x;y)] is equal to the information from the sender [H(x)] plus the information from the receiver [H(y)] minus the information in the joint occurrences [H(x,y)].[7,8,9] These quantities can be computed from the following formulae where n_{ij} is number of entries in the i,j location in the matrix:

$$H(x) = \log_2 N - \frac{1}{N} \sum_i \left(\sum_j n_{ij}\right) \log_2 \left(\sum_j n_{ij}\right)$$

$$H(y) = \log_2 N - \frac{1}{N} \sum_j \left(\sum_i n_{ij}\right) \log_2 \left(\sum_i n_{ij}\right)$$

$$H(x,y) = \log_2 N - \frac{1}{N} \sum_{ij} n_{ij} \log_2 n_{ij}$$

Where $N = \sum_{ij} n_{ij}$

$$T(x;y) = H(x) + H(y) - H(x,y)$$

These are the same formulae described by Metz et al in 1973, and applied to radiological images.[10]

The information transmitted may be calculated with the aid of an electronic calculator or with the aid of a table of logarithms to the base two. For the specific problem at hand the following values were calculated:[4]

	Technique I	Technique II
T(x;y)	1.39	0.79
Accuracy	89%	78%

These values suggest the superiority of Technique I over Technique II from both the accuracy and information transmitted metrics.

Conclusion

The evaluation of different medical imaging techniques is becoming more and more difficult because sometimes the changes in techniques are so minimal that standard evaluation procedures will not show a difference between techniques. In this situation, we must make use of all available data. Since medical imaging is not just a problem of detectability, and since in most cases a specific diagnosis will be suggested, procedures from information theory can provide a way of analyzing the multiple choice problem.

What remains to be done is to consider the stability of this parameter and how it affects the quality of medical care.

REFERENCES

1. D.J. Goodenough, K. Rossmann, and L.B. Lusted, "Radiographic Applications of Receiver Operating Characteristic (ROC) Curves," Radiology 110, 89(1974).

2. P. Rosenberg, "Resolution, Detectability and Recognizability," Photogammetric Engineering 37, 1255 (1977).

3. S.J. Starr, C.E. Metz, L.B. Lusted, and D.J. Goodenough, "Visual Detection and Localization of Radiographic Images," Radiology 116, 533(1975).

4. C.E. Metz, S.J. Starr, and L.B. Lusted, "Observer Performance in Detecting Multiple Radiographic Signals," Radiology 121, 337 (1976).

5. S. Sostre, J.D. Quinones, J.B. Schieve, J.M. Zimmerman, and A.B. Ashare, "Information Content of Thyroid Scintigrams: Pinhole versus Rectilinear Scanner," to be presented at the 63rd Annual Meetingof the Radiological Society of North America, Chicago, 1977.

6. D.A. Turner, "Scintillation Camera versus Rectilinear Scanner for Liver Imaging," J. Nuclear Medicine 17, 419 (1976).

7. F. Attneave, Applications of Information Theory to Psychology, Holt-Dryden, New York, 1959.

8. R.D. Luce, "Detection and Recognition," P. 103 in Handbook of Mathematical Psychology, R.D. Luce, R.R. Bush, and E. Galanter, Editors, Wiley, New York, 1963.

9. W.J. McGill, "Multivariate Information Transmission," Psychometrika 19, 97 (1954).

10. C.E. Metz, D.J. Goodenough, and K. Rossmann, "Evaluation of Receiver Operating Characteristic Curve Data in Terms of Information Theory, with Applications in Radiography." Radiology 109, 297 (1973).

Acknowledgments

I wish to acknowledge the following people: Dr. Samuel Sostre, Head of Nuclear Medicine at Buffalo General Hospital for suggesting the problem. Dr. Herbert Colle of the Department of Psychology at Wright State University for suggesting possible solutions and continuing collaboration. Dr. Charles Metz of the Department of Radiology at University of Chicago for helpful hints and reviewing the manuscript. Drs. Matthew Kabrisky and Joseph Carl of the Department of Electrical Engineering, Air Force Institute of Technology at Wright-Patterson Air Force Base for their help and guidance.

DISCUSSION:

Ashare noted that as far as he knew, no one was using the multiple category approach to medical decision-making.

Metz stated that the reason that people have so far been prevented from using a generalized ROC analysis for parallel multi-alternative decision-making is that there does not exist a single decision level or threshold; that is, more than one degree-of-freedom.

Ashare stated that one must look for other techniques, since people are using standard ROC techniques to compare two different tests and are not getting meaningful results.

Brill summarized by noting that ROC analysis is for signal detection and when one attempts to compare CT and Nuclear Medicine images, for example, one really wants to know which procedure characterizes the state of the patient best rather than just whether the study is normal or abnormal.

Metz agreed that other techniques should be explored but noted that some of these techniques may suffer from poorly determined prior probabilities.

A NEW METHOD OF EMISSION TOMOGRAPHY USING COMPOUND COLLIMATION

R.A. Vogel, D.L. Kirch, M.T. LeFree, W.C. Klingensmith, R.S. Trow, P.P. Steele

From the Division of Cardiology, Department of Medicine,
and the Nuclear Medicine Service, Department of Radiology
Denver Veterans Administration Hospital and the University
of Colorado Medical Center, Denver, Colorado

ABSTRACT

A new method of volumetric emission tomography is described using a wide-field Anger camera and a seven pinhole collimator. Each pinhole acquires data simultaneously from the emissive source projecting the data onto seven independent regions of the detector surface at different viewing angles. A mixed addition-multiplication reconstruction algorithm is used and multiple planes are reconstructed by altering the superpositional relationship among the back-projected image views. Myocardial perfusion studies using thallium-201 have demonstrated scintigram defects compatible with clinical information which were not visualized by non-tomographic radionuclide imaging techniques.

A NEW METHOD OF EMISSION TOMOGRAPHY

Introduction

In contrast to recent advances in computerized transmission axial tomography which have resulted in wide clinical application, emission tomography has shown only modest clinical usefulness. This is despite introduction of radionuclide tomographic scanning by Kuhl and Edwards 14 years ago[1,2] and tomography using an Anger camera technique described by Muehllehner[3] seven years ago. This report describes a new emission tomographic technique using a standard 37.5 cm wide-field Anger camera and a multi-sectorial collimator which views seven independent projections of a volume of interest simultaneously. Computer processing of clinical data by this method produces emission tomograms which are demonstrably superior to corresponding nontomographic scintigrams.

Methods

Description of collimator: As seen in Figures 1,2, and 3, a lead collimator was constructed with seven 5 mm diameter pinholes placed in a plane parallel to and 12.7 cm from the camera crystal. The six peripheral pinholes are 6.35 cm from both the center pinhole and from each other on a hexagonal pattern as shown in Figure 2. Figure 3 shows the lead shielding located between the pinhole plane and the crystal. This shielding prevents overlap of projected pinhole images. The central collimator pinhole has a conical field-of-view of 53° apical solid angle. The outerholes each have conical fields-of-view of 45°. The central conical field of view is perpendicular to the crystal face, whereas the peripheral conical fields' longitudinal axes converge inward at 26.5^{o}. Thus a conically tipped 12.7 cm diameter cylindrical volume is formed within which emissive data projects through all seven collimator holes independently to seven crystal sectors, each viewing the volume from a different spacial angle. The mean angle between views from opposite outside peripheral pinholes is 53° and the angle between the central view and each of the outer views is 26.5^{o}.

Emissive data within any plane parallel to the collimator face projects to the crystal as seven geometrically identical inverted images whose size relative to the emissive source is dependent upon the source's distance from the collimator. Additionally, data more distant along a line perpendicular to the collimator face project more centrally within the peripheral crystal sectors, whereas more proximal data project toward the outer edge of these crystal sectors.

Image reconstruction algorithm: As with any planar tomographic technique, data from the multiple projections within a particular plane of interest are superimposed whereas data out of the plane of interest project in a nonsuperimposing fashion. Each scintigraphic study was composed of an initial single projection data acquisition using the compound collimator. An Ohio-Nuclear wide-field Anger camera (model Sigma-410) was employed using a 20% energy window peaked high on the ascending spectral limb. Dynamic uniformity field correction using thallium-201 flooding was employed and image data was recorded on an Ohio-Nuclear model 75 tape recorder.

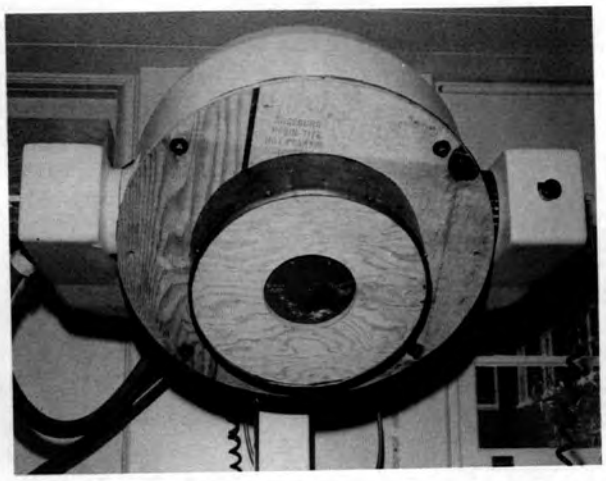


Figure 1. Ohio-Nuclear Anger camera (Model Sigma-410) with tomographic pinhole collimator in place. Plywood is used for structural rigidity.

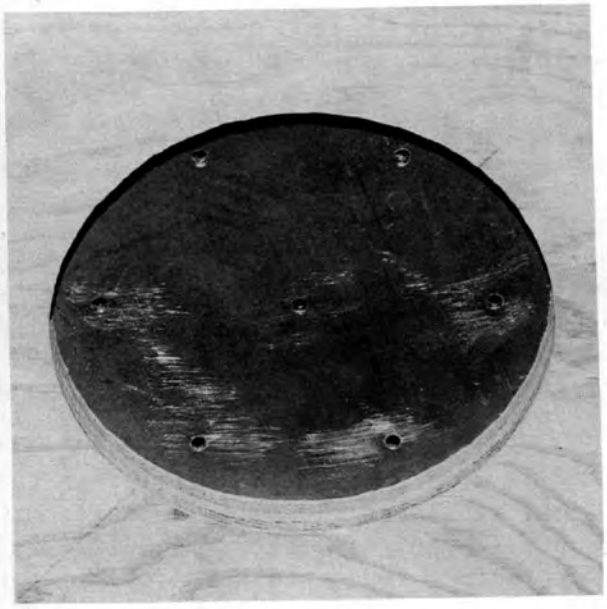


Figure 2. Close-up view showing detail of the seven pinhole array. Pinholes are 5 mm in diameter.

Figure 3. Inside of collimator head showing lead shielding used to prevent the images projected through each pinhole from overlapping.

Following the data acquisition, a second scintigram was obtained of an emissive point source (thallium-201) centered 12.7 cm from the collimator face. This provided a corresponding reference point projection within each of the seven views. Once known, all emissive data within the 12.7 cm plane can be superimposed by matching the location of the seven point source reference locations. In practice, using a Digital Equipment Corporation PDP-12 digital computer, the scintigraphic data within each of the seven crystal sectors was digitized centered about the location determined by the reference point projection and stored in 64 by 64 element array in disk memory. Opposite peripheral images were added in an element-by-element fashion and a composite image was calculated by taking the mathematical product of the three peripheral summed images and the central image.

A series of other planes at 4 mm separation were then reconstructed by shifting the relative correspondence of the peripheral images in an inward or outward axial fashion, inward shifting resulting in reconstruction of planes closer to the collimator and outward shifting producing more distant planes. The tomographic images from each study were then photographed from the computer image display.

Resolution Measurements:

Using the above technique, full-width-half-maximum (FWHM) resolution measurements were made using a line source (capillary tube) of thallium-201 located 12.7 cm from the collimator with 10 cm of tissue equivalent density scattering material interposed. Additionally, a standard 6,9, 12 and 25 mm parallel bar phantom was placed over a cobalt-57 flood source and imaged at 12.5 cm distance from the collimator.

Clinical Studies

Four myocardial perfusion scintigrams were performed using intravenous injection of 2 mCi of thallium-201 using both the compound collimation emission tomographic technique and standard planar scintigraphy (high-sensitivity collimator). Resting studies were obtained 15 minutes following thallium-201 injection in two patients, one of whom had no angiographic evidence of coronary artery occlusive disease, and another with isolated circumflex coronary artery occlusion verified by coronary angiography. Studies were performed with the collimator face inclined 40° left anterior oblique. Two exercise thallium-201 studies were also performed in patients with multivessel coronary artery disease who had previously been shown to have normal non-tomographic studies following exercise. Both patients were exercised until they experienced angina pectoris associated with the appearance of electrocardiographic ST segment depression > 1mm. Two mCi of thallium-201 was then injected and the patients exercised for an additional 45 seconds. Tomographic scintigrams were made in the 40° left anterior oblique projection 10 minutes later, following which anterior-posterior and 40° left anterior oblique projections using a high sensitivity parallel hole collimator were made. Tomographic and non-tomographic scintigrams were made using a 20% energy window centered asymmetrically high with dynamic uniform field correction. Scintigrams were comprised of 500,000 total counts.

Tomographic and standard planar scintigrams were also performed on a thyroid following injection of 5 mCi of technetium-99m pertechnetate. Three hundred thousand total count images were obtained, 5 minutes following IV injection and imaging was performed in the anterior-posterior projection.

Results

Full-width-half-maximum resolution was found to be 10.3 mm with the line source at 12.7 cm and 10 cm of scattering material interposed. Six mm bars spaced 6 mm apart could be easily distinguished.

Both of the two resting thallium-201 perfusion studies correlated well with clinical information, the study performed on a known normal individual appeared normal with the aortic valve ring clearly visible on the superior aspect of the left ventricle (Figure 4). Similarly, the resting study on the patient with left circumflex occlusion showed a well localized posterior left ventricular wall defect (Figure 5). Of note is the fact that no such defect was visible on the nontomographic study.

The two patients undergoing exercise thallium-201 scintigraphy were chosen because of previously normal nontomographic studies despite known coronary artery disease. One of the two (Figure 6) demonstrated significant anterior and inferior left ventricular wall hypoperfusion on the tomographic study despite a normal appearing nontomographic scintigram. The second exercise scintigram had a normal appearance as did the corresponding nontomographic scintigram (Figure 7). Using 2 mCi of thallium-201 approximately 500 counts per second were detected which was similar to that obtained using a high-sensitivity collimator with the non-cardiac field shielded by lead drape.

The tomographic thyroid study demonstrated a cold left upper pole thyroid nodule which was also well visualized in the nontomographic study (Figure 8).

Discussion

Although the presented compound collimation tomographic technique utilized a process conceptually similar to other tomographic methods, the use of a single wide-field Anger camera to obtain simultaneously independent angles of view has not been previously accomplished. The development of this technique was initiated by this laboratory's interest in cardiac imaging with the constraints of a small relatively spherical organ configuration, a significant amount of unwanted background (lung) emissive contribution, and rapid organ motion inherent in cardiac imaging. Either first passage or gated blood pool imaging are well suited to this technique because of the simultaneous nature of data acquisition for all seven views.

The choice of the pinhole method of projection was based upon its good resolution at 10 to 20 cm distance from the collimator face and the constraint of multi-angle viewing. An initial trial with a multi-sectorial parallel hole (foil) collimator gave poor results due to a much poorer resolution at that distance. A higher-resolution, multisectorial parallel hole collimator is being constructed, which would circumvent the nonuniform depth dependent size of projection inherent in pinhole imaging.

The choice of the specific reconstruction algorithm is based both on consideration of the nonuniform sensitivity of pinhole imaging which is partially balanced out by use of summed opposite images, and by empirical findings. Straight seven sector addition, multiplication and other mixed addition-multiplication algorithms were employed, with the presented method being found superior. The method does produce significant contrast enhancement due to the fourth power multiplication algorithm employed. This tends to magnify differences among the higher emissive count densities and to eliminate the lower count data from visualization.

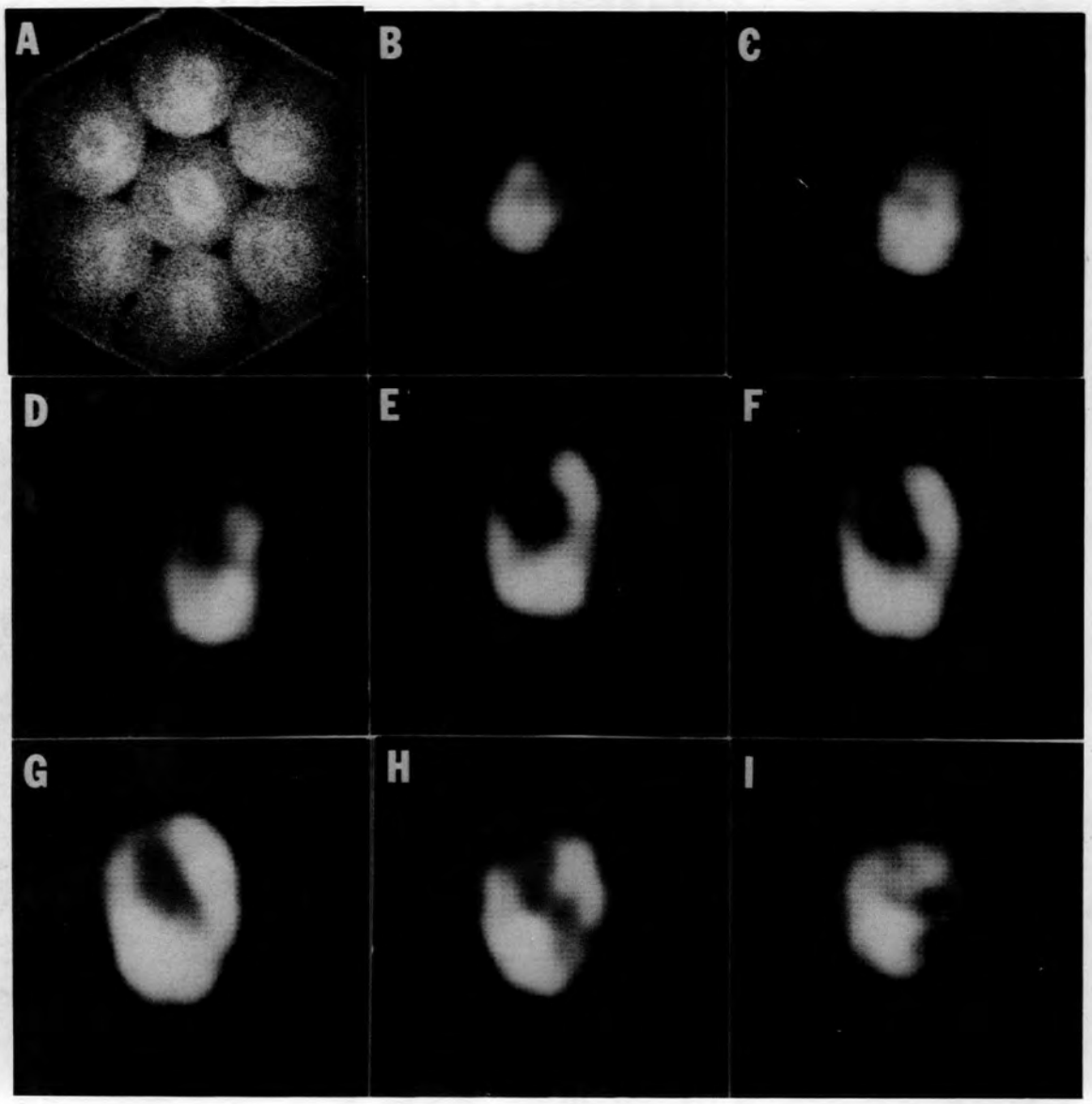


Figure 4. (A) Unprocessed tomographic thallium-201 scintigram of an angiographically normal individual. (B thru I) Tomographic images at 8 mm separation beginning at heart apex (B) and proceeding toward the mitral valve region (I).

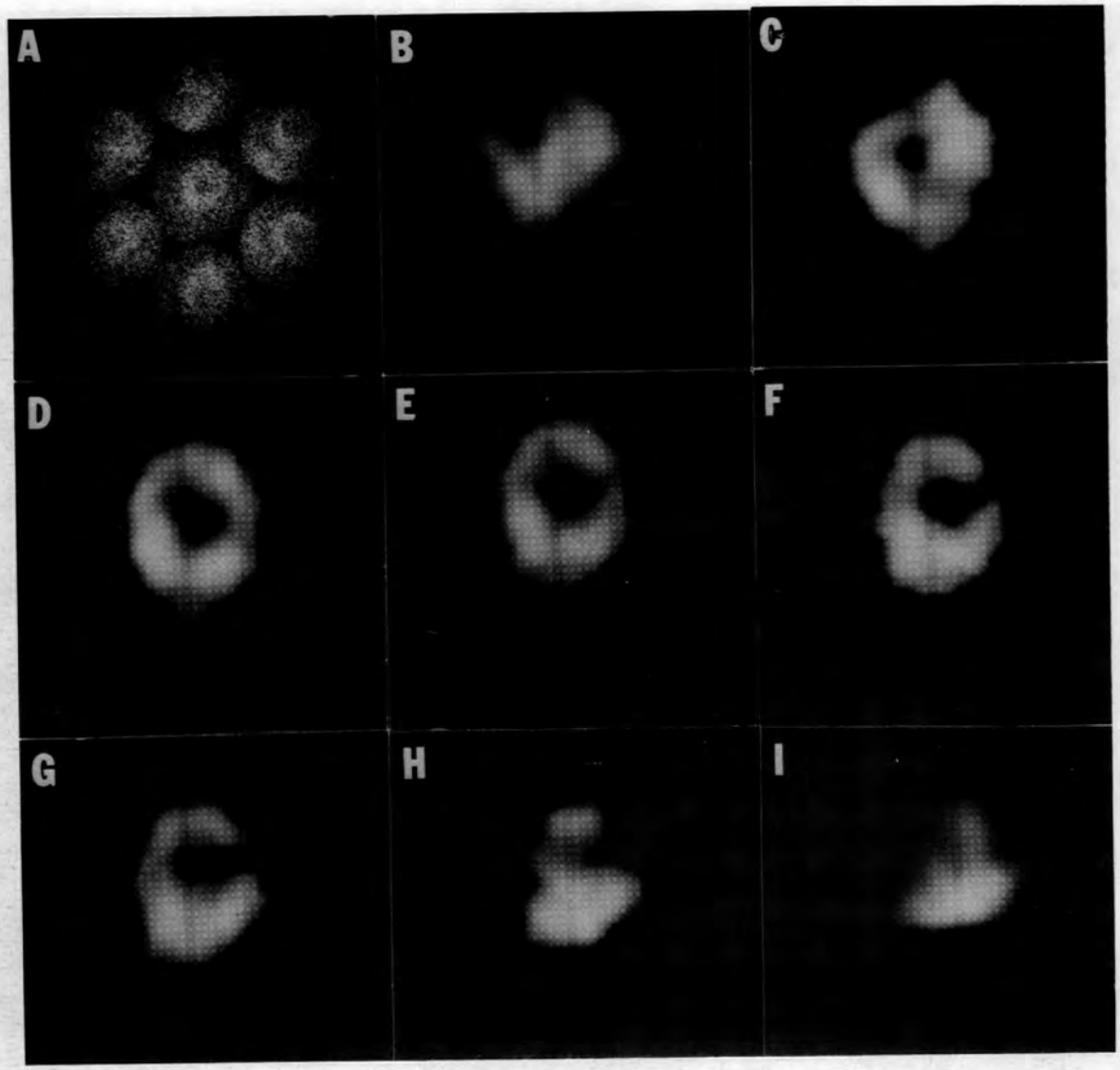


Figure 5. (A) Unprocessed tomographic thallium-201 scintigram of a patient with an angiographically demonstrated circumflex coronary artery occlusion. (B thru I) Tomographic images at 8 mm separation proceeding posterior from the cardiac apex (B) showing a posterior wall defect best visualized in (F) and (G).

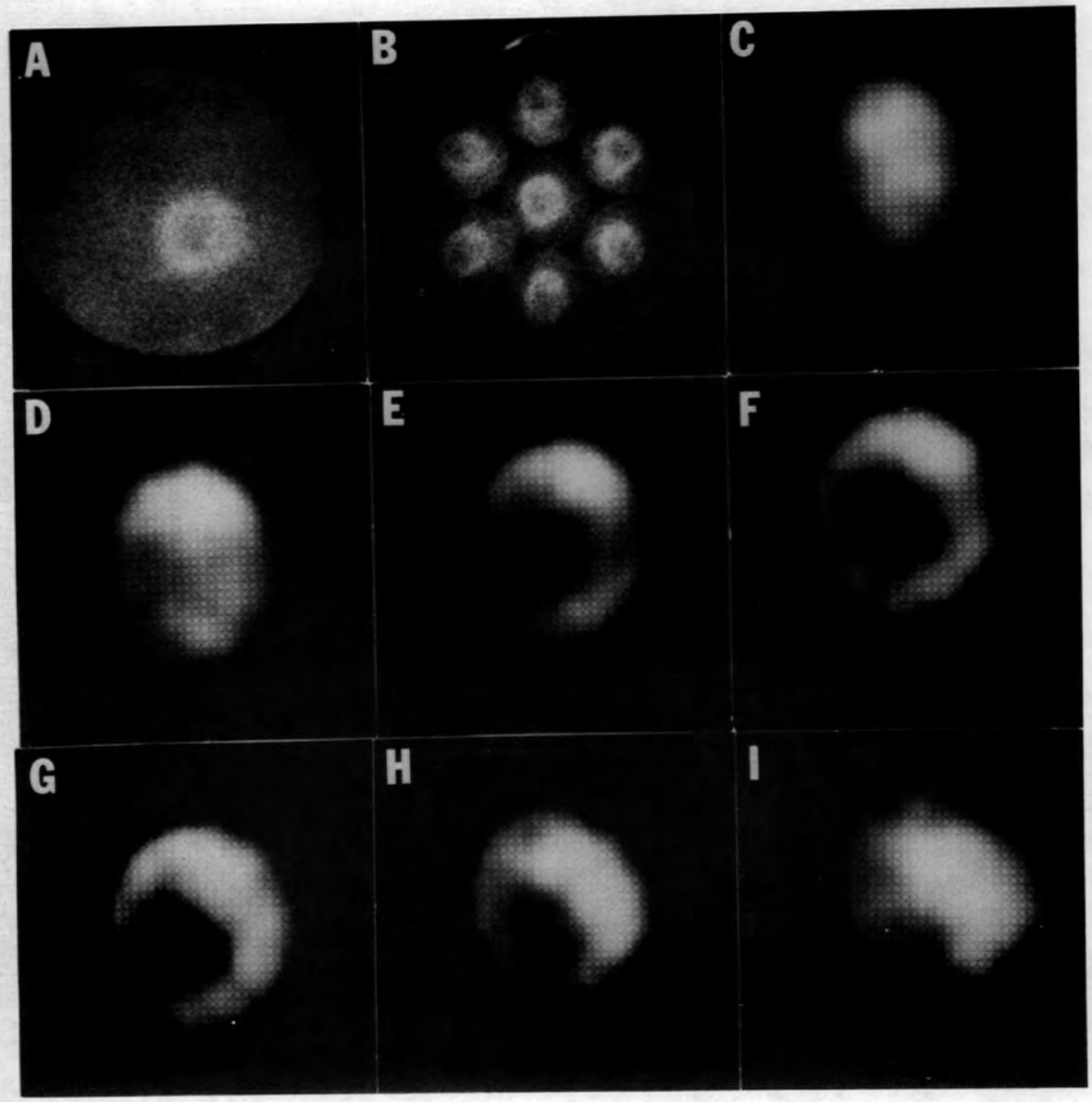


Figure 6. (A) Standard 40° LAO projection exercise thallium-201 scintigram of a patient with left anterior descending occlusion. (B) Unprocessed tomographic scintigram of the same individual. (C thru I) Tomographic images at 8 mm separation proceeding posteriorly from the cardiac apex (C) showing an anterior defect in the same individual.

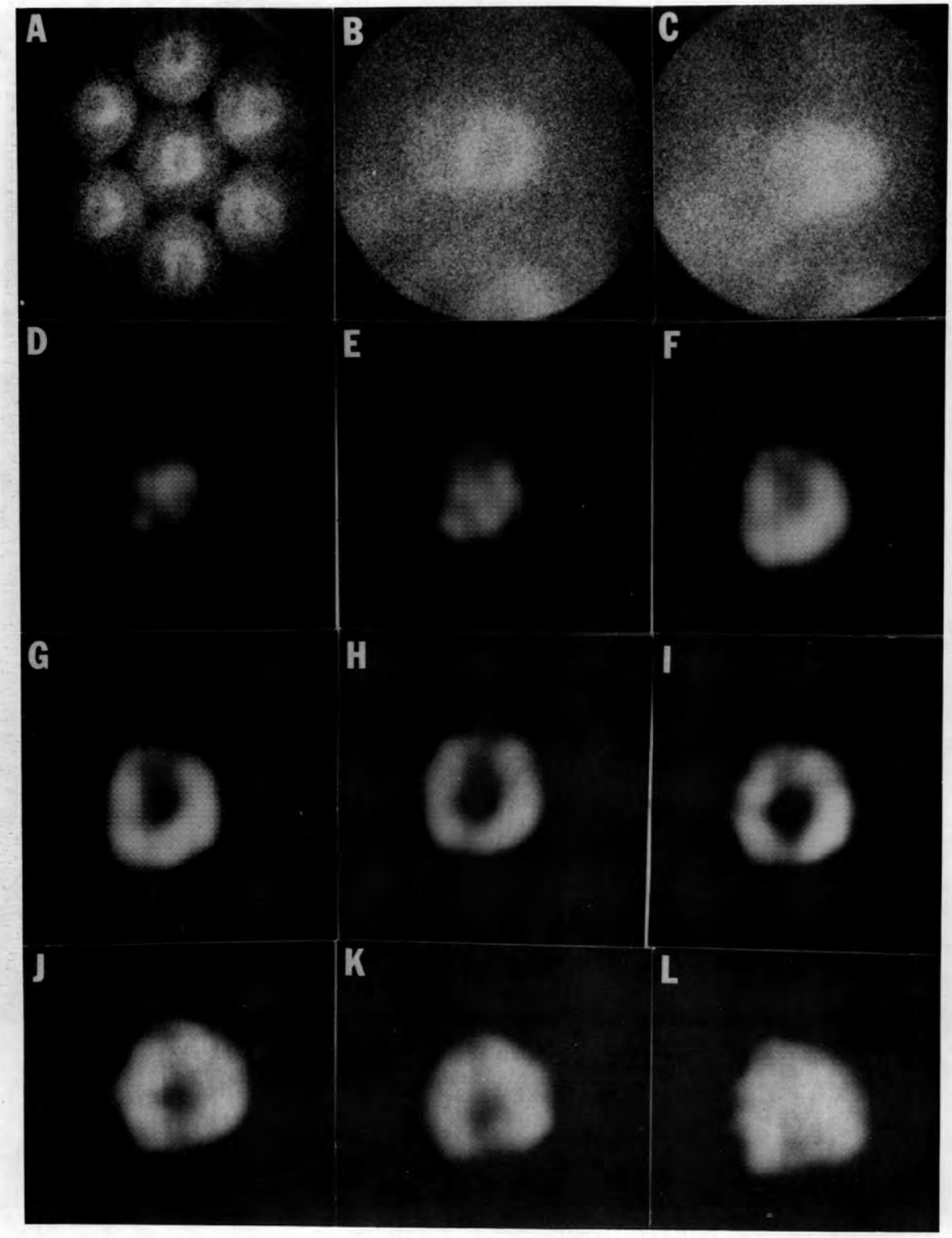


Figure 7. (A) Unprocessed tomographic thallium-201 scintigram of an individual with 3 vessel coronary artery occlusion. (B and C) 40° LAO and A-P projections respectively using a standard collimator. (D thru L) Tomographic images at 8 mm separation proceeding posteriorly from the cardiac apex (D) to the mitral valve (L).

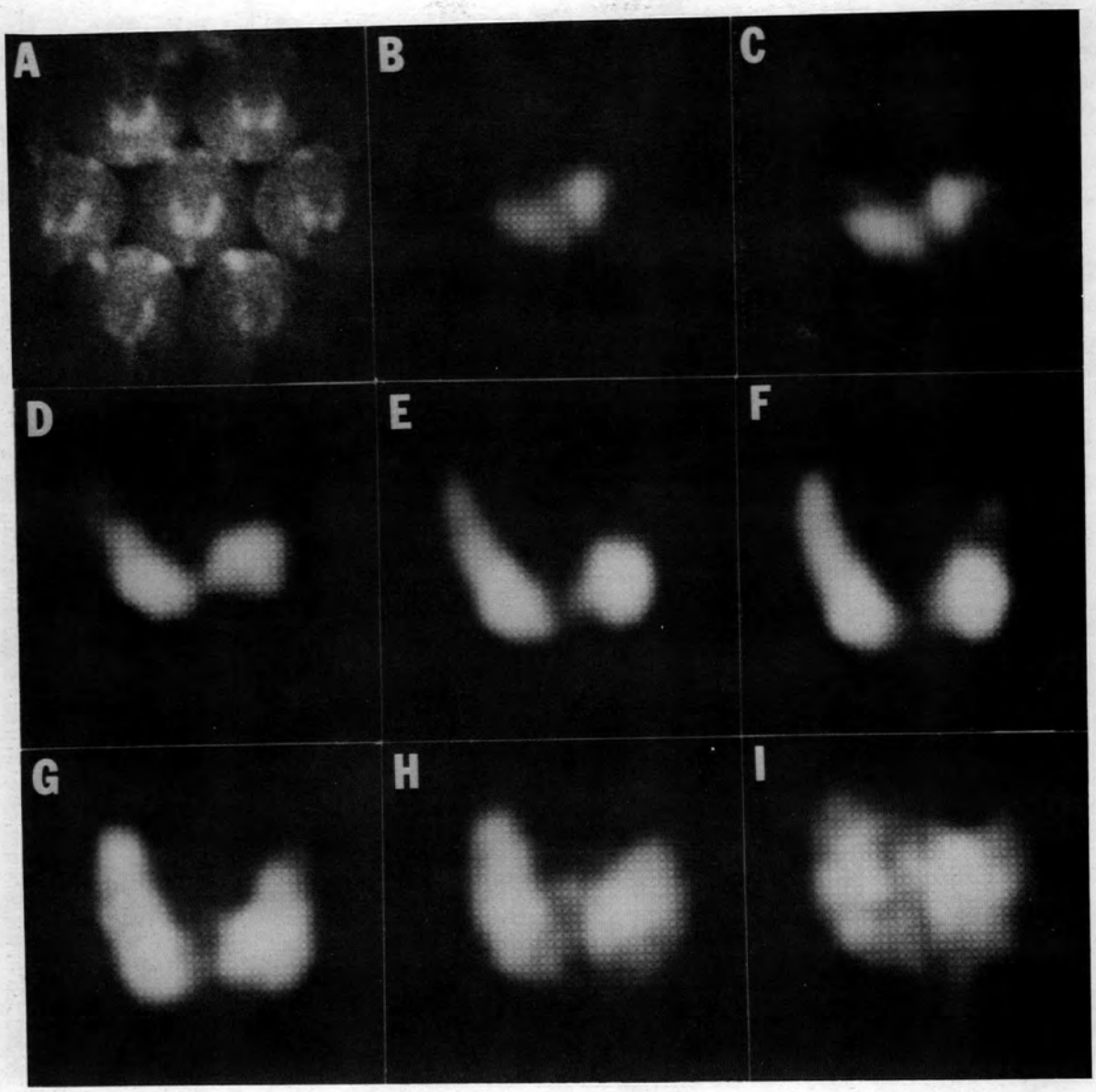


Figure 8. (A) Unprocessed tomographic thyroid scintigram of a patient with a cold left upper pole thyroid nodule using technetium-99m. (B thru I) Tomographic images at 4 mm separation of the same patient proceeding posteriorly from the thyroid isthmus (B).

In essence, however, the relevance of the clinical data obtained is the criterion on which methodological evaluation must be made. In this respect, the presented technique revealed data of appropriate clinical correlation not visualized by standard scintigraphy in two of the initial five studies performed. Despite improved acquisition resolution, no significant loss in organ-of-interest sensitivity was found. This initial report suggests that compound collimation emission tomography is both feasible and superior to nontomographic imaging of small organs.

REFERENCES

1. D.E. Kuhl and R.Q. Edwards, "Image Separation Radioisotope Scanning", Radiology 80, 653 (1963).

2. G.J. Hine and J.A. Sorenson, (editors), Instrumentation in Nuclear Medicine, Volume 2, Academic Press, New York & London, 1974.

3. G. Muehllehner, "Rotating Collimator Tomography", J Nucl Med 11, 347 (abstr) (1970).

MEASUREMENT OF LEFT VENTRICULAR VOLUME CHANGES WITH COMPUTERIZED RADIONUCLIDE ANGIOGRAPHY

J. Frischknecht, M. Quinlan, D. Kirch, M. LeFree, R. Vogel, P. Steele

From the Division of Cardiology, Department of Medicine, and the Nuclear Medicine Service, Department of Radiology Denver Veterans Administration Hospital and the University of Colorado Medical Center, Denver, Colorado

ABSTRACT

A radionuclide technique is described for obtaining high resolution, temporally resolved left ventricular volume curves. Technetium-99m, as pertechnetate, was injected into the superior vena cava and a time activity curve was extracted from an area of interest over the left ventricle with a single collimated scintillation probe. The collimation was then changed to view activity surrounding the left ventricle and a second injection was made to record background activity. The pair of curves generated by this method were recorded on magnetic tape for analytic refinement by a digital computer. The accuracy of the technique was tested by comparing left ventricular volume curves obtained from contrast ventriculograms and radionuclide ventriculograms in twenty patients undergoing diagnostic left heart catheterization using standard statistical methods. The average correlation coefficient obtained was .92 $\pm$ .01 ($\pm$ SEM). This method appears to be an accurate means of obtaining a temporally resolved left ventricular volume curve without left heart catheterization and is shown to be useful in the analysis of systolic ejection and diastolic pressure volume relationships.

MEASUREMENT OF LEFT VENTRICULAR VOLUME CHANGES WITH COMPUTERIZED RADIONUCLIDE ANGIOGRAPHY

Introduction

Attempts to measure cardiac volume in vivo began in 1914 with Geigel's method of deriving cardiac volume from a single x-ray picture (1). The development of angiocardiography provided a technique for visualizing the individual cardiac chambers. Gauer (2,3) and subsequently Rushmer and Thal (4) developed a technique for the evaluation of left ventricular volume. Methods for calculating left ventricular volume were introduced by Dodge and Tannenbaum (5) and Arvidson (6). Chapman (7) calculated left ventricular volume changes throughout the cardiac cycle. A single composite left ventricular volume curve was constructed by calculating left ventricular volume frame-by-frame from biplane cineangiograms at 30 frames per sec., related to filming time. Left ventricular volume changes related to changes in pressure and timing within the cardiac cycle were subsequently reported by Dodge, Sandler, et al (8,9). This approach was used in the study of mitral regurgitation by Khalaf et al (10) and the form of the volume curve was felt to be diagnostic of mitral regurgitation and inferred the degree of severity. Kennedy et al (11) used left ventricular volume time plots to compare volume obtained from biplane and single plane cineangiocardiography (A-P r=.88 and RAO r=.91).

Unfortunately angiocardiography requires left heart catheterization, pressure injection of large volumes of contrast and frame-by-frame calculation of volume is tedious. Left ventricular angiography has been reported to produce a significant increase in left ventricular volume beginning three to five beats after injection of contrast media (12) and is not, therefore, suitable for serial clinical studies at frequent intervals. Alternative non-invasive methods of measuring left ventricular volume include echocardiography, indicator dilution technique, radiopaque epicardial markers, radionuclide time-activity curves, radionuclide angiocardiography and gated scintiphotography. The only reported alternative methods of evaluating left ventricular volume changes throughout the cardiac cycle are by the use of radiopaque epicardial markers (13) and time-activity curves (14). Davila, et al, (15) have recorded left ventricular volume continuously by mechanically recording epicardial wall motion in dogs.

The purpose of this report is to describe a radionuclide method of obtaining a temporally resolved left ventricular volume curve by analysis of a pair of time-activity curves extracted from two regions of interest over the left ventricle with a scintillation probe using the collimation and injection technique reported in (16). The relationship between volume curves obtained from left ventricular cineangiograms, and radionuclide time-activity curves is tested by comparing these volume curves obtained sequentially in the same patient.

Methods

Twenty patients scheduled to undergo diagnostic left ventricular angio-

grams formed the study population. All patients were in sinus rhythm and gave informed consent. Catheterization was carried out in a fasting state after 5-10 mg of diazepam and 50 mg of diphenhydramine hydrochloride were administered orally. Retrograde left heart catheterization via brachial arteriotomy was carried out with a standard #7 french NIH catheter. Left ventricular cineangiograms were made by power injection of 40-60 cc's of contrast (Renograffin 76, Squibb) over four seconds in a 40° right anterior oblique projection filmed by a 35 mm airiflex-type camera at a speed of 60 frames/sec. Fifteen minutes were allowed for the depressant effects of contrast media on the left ventricle to resolve while the arteriotomy was repaired and preparation for recording the pair of radionuclide time-activity curves was made.

The midpoint of the left ventricle was estimated under fluoroscopy with the use of a radiopaque marker. The outer circumference of the probe collimation was outlined on the patient's chest with the midpoint of the left ventricle in the center of the collimator circle. A scintillation probe with a 5 cm x 5 cm sodium iodide crystal mounted on a portable stand, previously described (14,16), was positioned over the left ventricle by superimposition over the previously drawn circle.

Collimation for the first injection was provided by a 3.5 cm circular port through 1.3 cm thick lead. A catheter was placed in the superior vena cava and 5 mCi of ^{99m}Technetium sodium pertechnetate in 2.5-5.0 ml of solution was injected into the dead space of the catheter and flushed with 10 cc's of normal saline. The activity over the left ventricle was recorded with the probe perpendicular to the chest wall in an anterior position. This configuration of equipment and injectate produced count rates of 30 to 50 Kcps at the peak of the left ventricular phase of the study. Following the first injection the eclipse collimator described in (16) was placed on the probe and a second injection of 2.5 mCi was used to produce a background correction curve. Two representative curves obtained in this manner are shown in Figure 1 (A).

Since the objective of this study was to obtain high fidelity recording of the volume changes of the left ventricle as a function of time, simple paper strip chart recording techniques were inadequate. Therefore, the output of the analog meter from the single probe and a simultaneous electrocardiographic signal were recorded on a frequency modulated (FM) magnetic tape recorder for about 30 seconds following each injection. The band pass of the input and output amplifiers associated with this recording technique was 50 Hz. The radiocardiogram obtained was digitized at 100 samples/sec. on a PDP-12 digital computer where it could be analyzed and directly compared with the results of volume curves obtained by analyzing the contrast ventriculograms on the same patients.

DATA ANALYSIS

Left Ventricular Angiograms

The left ventricular angiograms were analyzed frame-by-frame through 2-4 successive cardiac cycles projected from a Tage Arno viewer. Beats were selected for analysis within the first 3-4 beats after contrast was injected and premature beats and beats immediately following a premature beat were discarded. Left ventricular volume was obtained by the area length method for single plane angiography as described by Kennedy et al (11) by the ellipsoidal formula: ($\pi LD^2/6$) where L is the measured length from apex to the midpoint of the aortic valve, D is the short axis derived from the longest length and the planimetered area (A) of the ventricle ($D = \frac{4A}{\pi L}$). The left ventricular angiograms were analyzed by a semiautomatic digital computer program which incorporates the above equation with an intractive,x-y digitizing

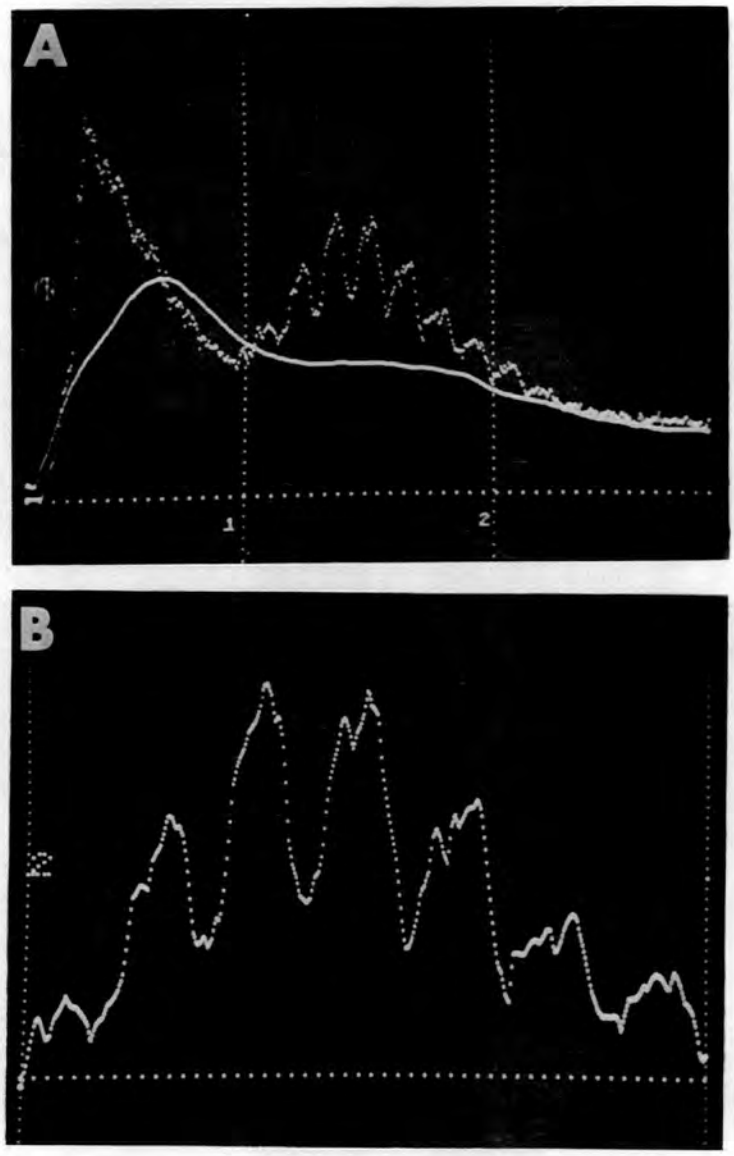


FIGURE 1: Illustration showing method for obtaining "flattened" LV volume curves from radionuclide single probe time-activity curves. In (A) the time-activity curve obtained with the probe collimated through the center of the LV shows beat-by-beat volumetric changes during the left ventricular phase of the study (between vertical cursors 1 and 2) and the smoothed background curve obtained with the eclipse collimator is displayed below. In (B) the background curve has been subtracted from the LV curve and the resultant corrected LV curve expanded between cursor 1 and 2 to fill the entire data buffer. It is then necessary to correct the LV volume data for tracer concentration variations as shown in Figure 2 (A). This "flattened" LV curve is obtained by dividing the data in (B) by a curve obtained by smoothing (B) to remove volumetric or beat-by-beat variations. The three beats nearest the peak would be selected for analysis.

table for tracing the outline of the left ventricle. Following the tracing of each outline and marking the long axis of the ventricle, the computer program computed the ventricular volume. Each value was stored on a magnetic disk for later comparison with the radionuclide ventriculogram.

Radionuclide Ventriculogram:

The analog radionuclide data was played back from the FM tape recorder and digitized at a rate of 100 samples per second. This data was stored for analysis by the technique illustrated in Figure 1. The left ventricular phase of the study is identified between the vertical lines (labelled 1 and 2) in Figure 1 (A). The curve obtained with the probe collimation viewing the center of the left ventricle (LV) shows the beat-by-beat volumetric variations expected. The smoothed background curve (BKGD) obtained with the probe collimation set to eclipse the left ventricle is shown below the left ventricular curve in Figure 1 (A). The method for obtaining this data and aligning these two curves for super-imposition has been described in detail in (16). These two curves are then subtracted (LV - DKGD) and the portion of the time axis between the vertical lines expanded to fill the full data buffer as shown in Figure 1 (B). This background corrected left ventricular time-activity curve has a mean value which is proportional to the concentration of the radionuclide tracer which exists during each beat. A curve which approximates this tracer concentration variation can be obtained by smoothing the curve in Figure 1 (B) to remove beat-by-beat volumetric fluctuations. This smoothed concentration curve is then divided into the curve shown in Figure 1 (B) to produce the "flattened " LV curve shown in Figure 2 (A). These "flattened" LV time-activity curves which are proportional to true left ventricular volume were then compared with LV volume curves obtained from frame-by-frame analysis of the contrast cineangiograms as shown in Figure 2 (A).

Comparison of Data

The LV volume curves obtained by the radionuclide and cineangiographic methods were expanded so that the same number of integral beats filled each respective data buffer. This was performed on three representative pairs of LV volume curves in Figure 2 (A). The upper curve in each comparison shows the data obtained by cineangiography and the lower curve shows the "flattened" radionuclide data in each instance. An EKG signal which was recorded during the radionuclide study is shown below each pair of LV volume curves. Prior to comparison of the two curves the data was biased so the mean values of two curves were equal. Following this normalization procedure the correlation coefficients shown below each set of curves was computed over the entire interval shown in Figure 2 (A). The LV ejection fraction was calculated from the contrast ventriculogram using end-diastolic volume (EDV) and end-systolic volume (ESV) for each beat. The ejection fraction ($EF = \frac{EDV-ESV}{EDV}$) was computed for each analyzable beat and the average computed for inclusion in Table 1.

Results

The patient's diagnosis, left ventricular ejection fraction from contrast cineangiogram and the correlation coefficient of the left ventricular cineangiogram and the radionuclide ventriculogram are summarized in Table 1. The average correlation coefficient of the ventricular volume curves for the twenty patients was $.92 \pm .01$ (SEM). As noted in Table 1 the poorest correlation occurred in those patients with the lowest ejection fractions. The data of patient 20 in Table 1 is displayed in Figure 2. Both curves are highly irregular and reflect an abnormal contraction pattern.

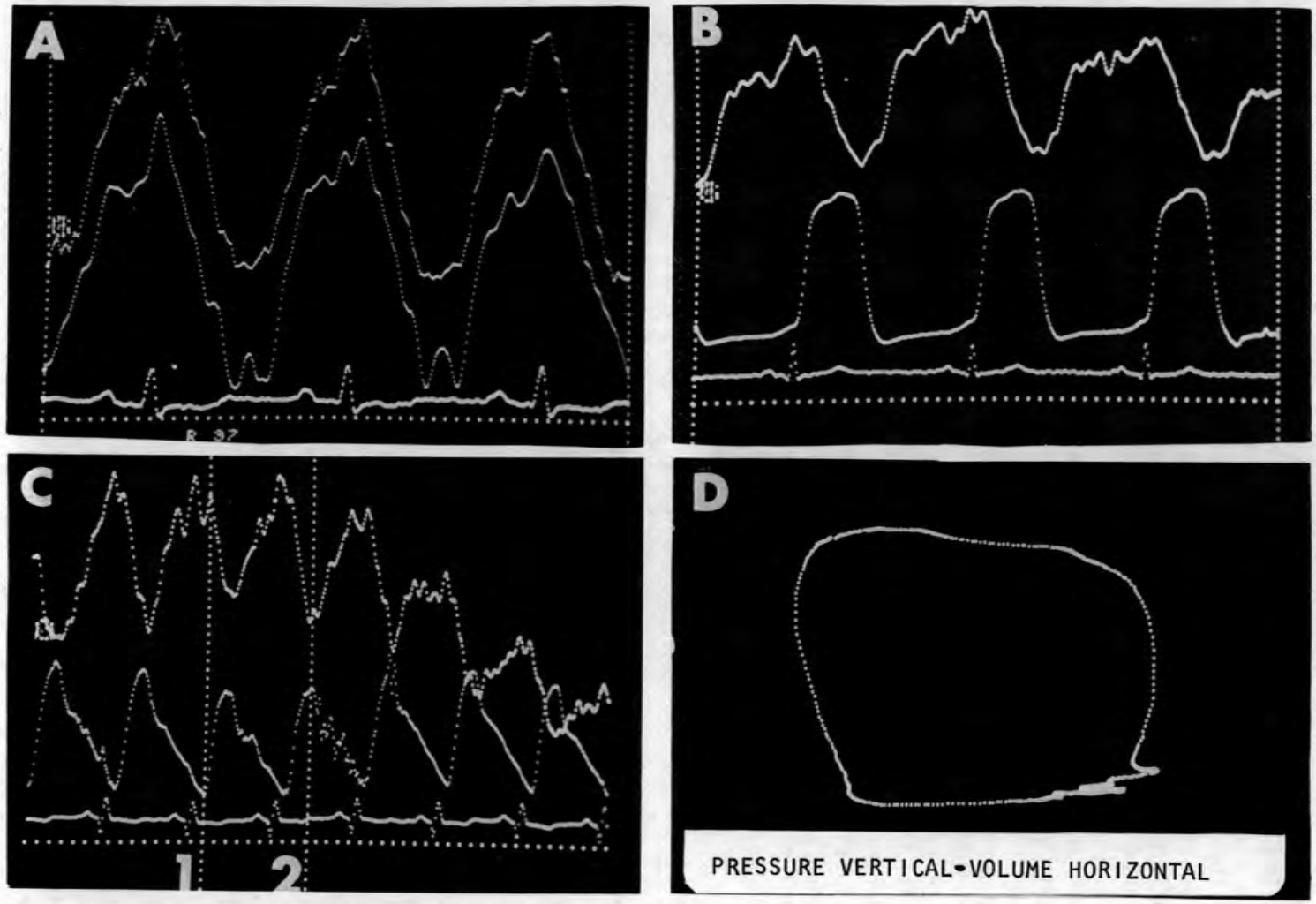


FIGURE 2: Three representative beats from patient 20 shown in Table 1 are shown in 2 (A). Cineangiogram (top curve) radionuclide ventriculogram (bottom curve) and EKG trace (baseline). Left ventricular and aortic pressure volume relationships are shown in (B) where simultaneous left ventricular volume (top), left ventricular micromanometer catheter pressure curve (middle) and EKG (bottom). In (D) the pressure volume loop from the third beat in (B) is displayed. Cursor 1 identifies the onset of systole and cursor 2 the end of systole. In (C) simultaneous left ventricular volume (top) and aortic micromanometer catheter pressure curve (middle) and EKG (bottom).

TABLE 1: SUMMARY OF PATIENT DATA

	DIAGNOSIS	LVEF CINE	CORRELATION COEFFICIENT, LV CINE, RADIONUCLIDE DATA
1.	Normal	.70	0.95
2.	Normal	.69	0.93
3.	Normal	.72	0.95
4.	Cardiomyopathy	.33	0.81
5.	CAD	.42	0.87
6.	CAD	.53	0.95
7.	CAD	.56	0.96
8.	CAD	.51	0.88
9.	CAD	.53	0.94
10.	CAD	.47	0.96
11.	CAD	.37	0.88
12.	CAD	.82	0.93
13.	CAD	.87	0.97
14.	CAD	.81	0.93
15.	CAD	.78	0.93
16.	CAD	.52	0.93
17.	CAD	.38	0.81
18.	CAD	.68	0.97
19.	MR	.60	0.96
20.	MR	.77	0.97
AVE ± SEM			0.92 ± 0.01

LEGEND: LVEF = Left ventricular ejection fraction; CAD = coronary artery disease; MR = mitral regurgitation.

Discussion

The method described in this report allows an accurate assessment of temporally resolved left ventricular volume changes. This method has the advantages of being noninvasive, risk free, suitable for frequently repeated studies, without affect of cardiac function, and nonirritating to the myocardium. The absorbed radiation dose is on the order of 2% of that received from a comparable angiographic study (14). Furthermore, changes in count rate during ventricular filling and contraction are directly related to volume changes rather than area changes which must be converted to volume when using contrast ventriculograms. Temporally resolved volume curves obtained in this manner reflect overall volume changes and are not suitable for regional wall analysis. Regional wall analysis has been accomplished with the use of dynamic radionuclide angiocardiography (17).

Clinical Implications

This method is potentially useful in the study of temporally resolved ejection and diastolic pressure-volume relationships. Ejection phase indices sum or average and values during systole and do not focus on systolic events. Despite severe coronary artery disease some patients have normal ejection phase indices. In an effort to reconcile coronary artery anatomy with function Johnson et al (18) have studied the rate of systolic ejection in three equal parts of systole. Normal ventricles have been shown to eject most of the blood during the first third of systole as opposed to ventricles with coronary artery disease and normal ejection fractions which eject most of the blood during the middle third of systole. Leighton et al (19) have analyzed segmental wall motion at midsystole in patients with coronary artery disease and normal ejection fractions and found decreased or paradoxical wall motion during the first half of systole which normalized by end systole. This observation in patients with normal ejection indices and abnormal anatomy necessitates analysis of systolic events. The correct identification of the onset and end of systole is imperative for the analysis of systolic events (21). By simultaneously recording ascending aortic pressure obtained by a micromanometer catheter and a temporally resolved radionuclide ventriculogram, the onset and end of ejection has been correctly identified. Figure 2 (C) illustrates that the onset of systole (cursor 1) as manifested by the generation of positive pressure in the aorta corresponds to the peak of the radionuclide ventriculogram. The end of systole (cursor 2) manifested by the peak of the aortic pressure corresponds to the first plateau of the radionuclide ventriculogram at the end of ejection. Temporally resolved radionuclide ventriculograms allows for analysis of systolic events and would be potentially useful in this type of study.

Diastolic stiffness of the left ventricle is a major determinant of left ventricular filling pressure and chamber size (20). The study of left ventricular compliance in man has encountered technical difficulties associated with simultaneous measurement of pressure and volume. The curvilinear nature of the pressure volume relationship necessitates pressure and volume analysis throughout diastole rather than relying on only one measurement at end diastole. Simultaneous echocardiographically determined volume and left ventricular pressure measurements reported by Grossman et al (21) is the only method available for simultaneous pressure volume measurements. Temporally resolved radionuclide ventriculograms allow for simultaneous volume and pressure measurements throughout diastole. Figure 2 (B) illustrates simultaneous left ventricular pressure and a temporally resolved radionuclide ventriculogram. Figure 2 (D) illustrates the pressure volume loop from the third beat shown in Figure 2 (B). Segmental wall abnormalities of the left ventricle are a potential source of error for echocardiographic volume studies

(22). Since our method measures volume directly segmental wall abnormalities are of no concern. Alderman et al (23) have shown that pressure-volume curves shift in response to experimental intervention with angiotension and nitroprusside. Radionuclide ventriculography can be repeated frequently without altering ventricular function and is a potentially useful method for measuring diastolic volume changes with experimental and clinical interventions. In conclusion a temporally resolved radionuclide ventriculogram is a safe, reliable method for obtaining high resolution volume curves of the left ventricle which does not alter contractility, or irritate the myocardium, can be repeated at frequent intervals and is potentially useful for the study of systolic and diastolic events as previously discussed.

REFERENCES

1. Geigel R: Die klinische Verwertung der Herzsilhouette. Munchen med. Wchnschr., 61: 1220, 1914.

2. Gauer OH: German aviation medicine. World War II p. 565, Washington, Dept. of Air Force 1950.

3. Gauer OH: Volume changes of the left ventricle during blood pooling and exercise in the intact animal: Their effects on left ventricular performance. Physiol Rev 35: 143-155, 1953.

4. Rushmer RF, Thal N: Mechanics of ventricular contraction; cinefluorographic study. Circ 4: 219-228, 1951.

5. Dodge HT, Tannenbaum HC: Left ventricular volume in normal man and alterations with disease. Circ 14: 927, 1956.

6. Arvidsson H: Angiographic observations in mitral valve disease with special reference to the volume variation in the left atrium. Acta Radiol Suppl 158, 1958.

7. Chapman CB, Baker O, Reynolds J, Bonte FJ: Use of biplane cinefluorography for measurement of ventricular volume. Circ 18: 1105-1117, 1958.

8. Dodge HT, Hay RE, Sandler H: An angiographic method for directly determining left ventricular stroke volume in man. Circ Res 11: 739-748, 1962.

9. Sandler H, Dodge HT, Hay RE, Rackley CE: Quantitation of valvular insufficiency in man by angiocardiography. Am Heart J 65: 501-513,1963.

10. Khalaf J, Chapman CB, Ernst R: The cinefluorographic approach to the diagnosis of mitral regurgitation. Prog Cardiovasc Dis 5: 230-247, 1962.

11. Kennedy JW, Trenholme SE, Kasser IS: Left ventricular volume and mass from single-plane cineangiocardiogram. A comparison of antero-posterior and right anterior oblique methods. Am Heart J 80: 343-352, 1970.

12. Carlton RA: Change in left ventricular volume during angiocardiography. Am J Cardiol 27: 460-463, 1971.

13. Vine DL, Dodge HT, Frimer M, Steward DK, Caldwell J: Quantitative measurement of left ventricular volumes in man from radiopaque epicardial markers. Am J Cardiol 54: 391-399, 1976.

14. Van Dyke D, Anger HO, Sullivan RW, Vetter WR, Yano Y, Parker HC: Cardiac evaluation from radioisotope dynamics. J Nucl Med 13: 585-592, 1972.

15. Davila JC, Sanmarco ME, Phillips CM: Continuous measurement of left ventricular volume in the dog. I description and validation of a method employing direct external dimensions. Am J Cardiol 18: 574-583, 1966.

16. Steele PP, Van Dyke D, Trow RS, Anger HO, Davies H: Simple and safe bedside method for serial measurement of left ventricular ejection fraction, cardiac output and pulmonary blood volume. B Heart J 36: 122-131, 1974.

17. Kirch DL, Steele PP, LeFree MT, Evans PC, Stern DM, Vogel RA: Application of a computerized image intensifier radionuclide imaging system to the study of regional left ventricular dysfunction. IEEE Transaction on Nuclear Science. Vol NS-23 No. 1, 1976.

18. Johnson LL, Ellis K, Schmidt D, Weiss MB, Cannon PJ: Volume ejected in early systole. Circ 52: 378-389, 1975.

19. Leighton RF, Pollack MEM, Welch TG: Abnormal left ventricular wall motion at midejection in patients with coronary heart disease. Circ 52: 238-244, 1975.

20. Levine HJ: Compliance of the left ventricle. Circ 46: 423-426, 1972.

21. Grossman W, Stefadouras A, McLaurin LP, Rolett EL, Young DT: Quantitative assessment of left ventricular diastolic stiffness in man. Circ 47: 567-574, 1973.

22. Ludbrook P, Karliner JS, Peterson K, Leopold G, O'Rourke RA: Comparison of ultrasound and cineangiographic measurements of left ventricular performance in patients with and without wall motion abnormalities. Br Heart J 35: 1026-1032, 1973.

23. Alderman EL, Glantz SA: Acute hemodynamic interventions shift the diastolic pressure-volume curve in man. Circ 54: 662-671, 1976.

INTERNAL DISTRIBUTION

1. L. S. Abbott
2. L. L. Anthony
3. J. A. Auxier
4. A. A. Brooks
5. H. P. Carter
6. H. Goldstein (Consultant)
7. S. V. Kaye
8. F. C. Maienschein
9. W. J. McClain
10. B. F. Maskewitz
11. J. K. Poggenburg, Jr.
12. H. Postma
13. C. R. Richmond
14. A. Zucker

15-82. BCTIC Distribution

83-84. Central Research Library

85. ORNL Y-12 Technical Library Document Reference Section

86. Biology Division Library

87. Laboratory Records Dept.

88. Laboratory Records R.C.

89. ORNL Patent Office

EXTERNAL DISTRIBUTION

90. T. M. Albert, Office of Environment and Safety, U.S. Department of Energy, Washington, D.C. 20545
91. William H. Blahd, M.D., VA Wadsworth Hospital Center, Nuclear Medicine SVC (691/115), Los Angeles, California 90073
92. J. Gitlin, Div. of Med. Rad. Exposure, FDA Bureau of Radiological Health, Rockville, Maryland 20852
93. M. Glos, Society of Nuclear Medicine, 475 Park Ave., So., New York, N.Y. 10016
94. J. L. Liverman, Div. of Biom. and Env. Research, U.S. Department of Energy, Washington, D.C. 20545
95. A. B. McIntyre, FDA Bureau of Radiological Health, 5600 Fishers Ln., Derwood, Maryland 20855
96. P. Paras, FDA Bureau of Radiological Health, 5600 Fishers Ln., Derwood, Md. 20855
97. W. Weyzen, Human Health Studies Programs, Div. of Biom. and Env. Research, U.S. Department of Energy, Washington, D.C. 20545

98-99. Technical Information Center, U.S. Department of Energy, Oak Ridge, Tenn.

100. Research and Technical Support Division, ORO, U.S. Department of Energy, Oak Ridge, Tenn.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1977-748-189/365