Image processing: Mathematics, engineering, or art?

Kenneth M. Hanson

Los Alamos National Laboratory
Mail Stop P940, Los Alamos, New Mexico 87545

Abstract

From the strict mathematical viewpoint, it is impossible to fully achieve the goal of digital image processing, which is to determine an unknown function of two dimensions from a finite number of discrete measurements linearly related to it. However, the necessity to display image data in a form that is visually useful to an observer supersedes such mathematically correct admonitions. Engineering defines the technological limits of what kind of image processing can be done and how the resulting image can be displayed. The appeal and usefulness of the final image to the human eye pertains to aesthetics. Effective image processing necessitates unification of mathematical theory, practical implementation, and artistic display.

Introduction

Figure 1 shows the basic elements of any imaging scheme. The fundamental purpose of imaging is to convey information about the object to the observer, usually a human being. The measurements obtained at the input stage of imaging can assume various forms. They might consist of spatially separated samples of the luminosity of visibly detectable light, as in light photography. Or, as is most often the case in medical imaging, the measurements might be of nonvisual quantities, such as x-ray intensity, the strength or time delay of sonic pulses, or the intensity of radiation being emitted by the object. In the newest form of medical imaging, that of nuclear magnetic resonance, the measurements involve a complex arrangement of magnetic and radiofrequency fields and the quantities being imaged are closely related to the density of the nuclei under study in combination with the relaxation times of the nuclear spins. Between the measurements and the display of the final image, some form of processing takes place. In photography or film-based radiography, the processing consists in film development. We will be more concerned here with digital image processing in which the measurements are manipulated by a digital computer. In order to emphasize the unity of the processing and display stages of imaging, we will assume the term "image processing" comprises both. It is typically desired that the observer synthesize the displayed information in order to draw a conclusion (make a diagnosis) about the object. Thus, the available information should be presented to the observer in such a way that he can most readily interpret it. Presently, the most efficient way to present the human observer with a vast amount of correlated information is through his visual sense. Thus, we will assume the end product of image processing is a visual image or picture. Indeed, those who practice image processing are

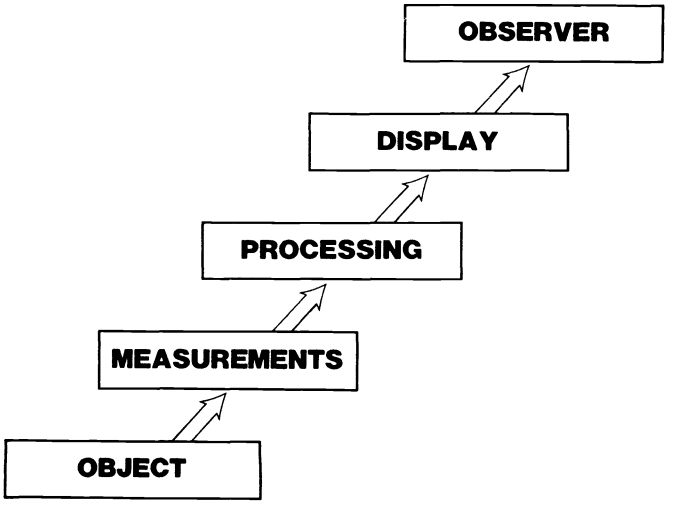


Figure 1. It is the purpose of imaging to provide an observer, usually a human being, information about the object under study. The information content is fundamentally limited by the measurements taken. Image processing, which naturally includes both the processing and display stages of the imaging chain, should convey as much of that information as possible in a form useful to that observer.

fortunate to have something visible to show for their efforts. This is one of the most appealing aspects of image processing. However, sight of the final image is too often lost in the euphoria that accompanies the comprehension of the mathematics. While it is hard to imagine, it may be possible in the future to transmit information to the brain in an entirely different way, which could prove more effective than vision.

The foundation of image processing rests on mathematics. Mathematical theory assures us that it is impossible to unambiguously determine an arbitrary function of two dimensions from a finite number of discrete, noiseless measurements linearly related to it. This impossibility is summarized under the concept of the null space of functions associated with any measurement scheme.[1,2] Thus, it would appear to follow that from a finite number of measurements, no reasonable estimate of the original image can be inferred.[3] In view of all the successful applications of image processing existent, such a conclusion is ridiculous, of course. Those who see only the mathematics are blind to the objective, the final image. In this paper we will discuss the roots of the above-stated mathematical restrictions, how they are overcome to produce a useful image and some of the factors that affect the usefulness of the image for human visual consumption. The underlying principle is that the goal of image processing is the production of a useful final image.

As we shall see, the displayed results relate fundamentally back to the initial measurements. While the design of the measurement scheme should be considered in a unified approach to any imaging problem, we will not discuss such design in any detail here. In medical imaging, diagnosis is more often made on the basis of patterns discerned in the images than on the basis of absolute image values. Thus, quantitative imaging will not be addressed, even though it may be useful in other contexts. Many of the examples used to demonstrate the ideas presented are related to computed tomography (CT). This is mostly because the author has had extensive experience in this field. However, the unusual incongruity between the measurement geometry and the normal display geometry inherent in CT makes it a provocative modality in which to learn image processing concepts.

Mathematical Foundations

We consider an imaging situation in which it is desired to determine and display a quantity $f(x,y)$ that is a function of the two continuous variables x and y (usually spatial coordinates). The quantity f may be some physical variable containing information about the object under study. For simplicity, let us assume the measurements are linearly related to f. The ith discrete measurement may be written as

$$g_i = \iint h_i(x,y)f(x,y)dxdy \qquad (1)$$

where h_i is the response function, or weighting function, that describes how much the value of f at each point (x,y) contributes to the ith measurement. The objective of image processing is to reconstruct or restore the function f from the given g_i and present it to the observer. The author has described in previous work[2,1,4] the interpretation of the functions f and h as vectors in a Hilbert space and the implied consequences concerning the inversion of Eq. (1). Rather than repeat those incantations here, let us consider a simpler, but less complete, approach. Suppose the measurements are actually projections or line integrals, as in CT. Then each h_i is a 2-D δ-function; zero everywhere except on the straight line of integration. Figure 2 shows the lines of integration that might be available for a coarsely-sampled CT measurement scheme. Clearly, the functional values of f in the regions between the lines do not contribute to the measurements. Because the data carry no information about f in the regions between the lines, the values of f in these regions cannot be reconstructed from the data. This inability to determine certain aspects of f corresponds to the existence of a subspace in the Hilbert space of f known as the null space. A similar manifestation of the null space is the situation in which one is presented with a photograph of El Tovar Lodge. It would be impossible to infer what a photograph taken at right angles to the first one would look like, unless one knew beforehand that El Tovar is perched on the rim of the magnificent Grand Canyon. Now suppose the measurements consist of strip integrals instead of line integrals, such that each strip is centered on the formerly used line and is just wide enough to touch the neighboring strip. Then the measurements suggested by Fig. 2 would completely cover the area shown. The value of f at each point would contribute to one and only one strip integral from each direction. It should not be surprising that in such a limited measurement geometry, a significant null space still exists, even though the previous argument falls down.

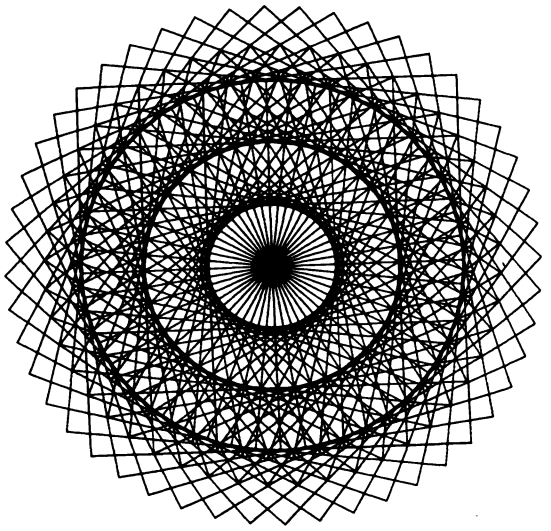


Figure 2. The ray paths that correspond to coarsely sampled projection measurements (24 views, each containing 7 samples). If the measurements truly consist of line integrals along the lines shown, the regions that lie between the lines do not contribute to the available measurements and hence correspond to the null space.

The presence of a null space is essentially a result of a many-to-one transformation, which inherently cannot be inverted without ambiguities. A finite number of discrete measurements of a function of continuous variables, as are used in digital image processing, comprise an infinite-to-one transformation. Hence, the original function can never be unambiguously recovered. Further, since all physical measurements are subject to finite resolution, they always have a null space associated with them. Therefore, all imaging systems have a null space corresponding to the lack of information at arbitrarily high spatial frequencies. The existence of the null space associated with the limited-data problem has been known for a long time.[3,5-15] However, its explicit effect in practical reconstruction problems has not been well elucidated. A simple method of generating the null-space part of an arbitrary object function corresponding to any CT measurement geometry is presented in Ref. 1. The author has found the concepts of Hilbert space and of the null space associated with measurements/transformations to be extremely useful in approaching a variety of problems in image analysis. For example, the null space accounts for the following: the renowned aliasing effect that accompanies discrete sampling[16]; the artifacts attendant with the restoration of blurred images; the limitations inherent in limited-angle tomography[1], and, in general, the infinity of solutions to ill-posed problems.[5]

The measurements are blind to the null space; they provide no information about the components of any original function f lying in the null space. That part of f not contained in the null space can be determined from the measurements. The corresponding subspace, which is orthogonal to the null space, is appropriately called the measurement space. Since the measurement space is comprised of all functions that can be written as a linear combination of the response functions, the expansion

$$\hat{f}(x,y) = \sum_i a_i h_i(x,y) \qquad (2)$$

provides a means for constructing an estimate that is wholly contained in the measurement space. This expansion is so instinctive that the response functions have been called "natural pixels."[17] As an aside, the dimension of the measurement space can be found by performing an eigenanalysis of the gramian matrix, the elements of which are the overlap integrals between each pair of response functions. The number of nonzero eigenvalues gives the dimensionality. This has been called the "number of degrees of freedom"[18] of the measurement geometry, and, in some sense, indicates the number of independent pieces of information that the measurements possess. Of course, the distinction between a zero eigenvalue and a very small one is only possible (and perhaps meaningful) in pure mathematics. In applications where real, noisy data must be used, this distinction fades. Thus the number of degrees of freedom may not be a precisely defined number in practice. In fact, it is reasonable to soften the language of pure mathematics when dealing with real problems and say, for example, that a particular subspace is "nearly" complete, or that a matrix is "essentially" singular (non-singular, but with very small singular values), etc.

The ambiguities associated with the null space typically result in artifacts in reconstructions. These artifacts depend upon how the null-space components of the estimated function are handled.[1] In an attempt to make the inversion unique, which is a predilection of mathematicians, it is common practice to require that out of the many possible solutions, the particular one with minimum norm be chosen. This amounts to demanding the null-space components of the solution have zero amplitude. The minimum-norm criterion limits the reconstruction to have the form of the response-function expansion, Eq. (2). As in the El Tovar example above, a better guess of the null-space components can be made if there is prior knowledge available about the object under investigation. The Bayesian approach,[1,4,19,20] which entails the use of such prior knowledge, can sometimes reduce reconstruction artifacts when the prior knowledge is restrictive enough. In such cases, the null-space components of the reconstruction are replaced by a better estimate than zero through the use of the prior information. The types of prior knowledge that have been studied in conjunction with CT include non-negativity of f, known region of support of f, and structural information about object.[1,4] Somewhat akin to the minimum-norm condition in its action, the requirement of maximum entropy has been applied to image inversion.[21-27] Maximum entropy also intrinsically imposes a nonnegativity constraint on f, which may account for its improved performance over traditional reconstruction methods.[27] The maximum-entropy approach has been fervently espoused by some[1,27,28] as a fundamental principle. Others, while acknowledging the successful results of maximum-entropy practitioners, have found difficulty accepting the fundamental tenet that the value of the reconstruction at each location should be interpreted as a probability.[29] It is refreshing to hear one of the maximum entropists[30] admit there is no reason to prefer the maximum-entropy result over a host of other solutions; it is useful because it works (produces visually pleasing results). Note that when nonlinear constraints, such as nonnegativity or maximum entropy, are invoked, the concept of a Hilbert space is strictly no longer applicable, since the definition of a Hilbert space includes linearity. One must talk of the null set instead of the null space, etc.

The Observer

The human observer is the final link in the imaging chain. Because the objective of image processing is to present the observer with a displayed image that will allow him to draw maximal information about the object, it is important to understand some of the characteristics of the human observer. Ignoring color and motion, some of the aspects of the human eye-brain system worth considering are the following:

a) visual acuity (resolution)
b) threshold for detection of low-contrast signals
c) influence of display brightness
d) influence of surround brightness
e) Mach-band phenomenon and other illusions
f) tolerance for visual noise
g) ability to synthesize correlated patterns
h) ability to assess statistical reliability
i) ability to use prior information in interpretation

These observer characteristics range from the obvious to the more subtle. Those appearing at the top of the list probably spring to everyone's mind. It is obviously essential that the displayed image be large enough, have enough contrast and brightness, and be presented in a suitably lit environment.[31,32] The effect of random image noise on the ability of human observers to detect simple, low-contrast signals against a constant background has recently been studied extensively by Burgess et al.[33,34] They find humans can "noise average" nearly as well as a mathematically ideal observer. Furthermore, the human can do well over a wide range of display contrasts.[35] However, when visual noise becomes too severe, observer performance suffers. Although little is known about the influence of the Mach-band effect upon the interpretation of images, it may be non-negligible. It is worth remembering that although seeing may be believing, it may not represent the truth.

The last three items listed above deal with the higher level processing that the human brain can obviously perform. The radiologist is distinguished from the proverbial "trained observer" by his ability to use prior knowledge. Through his training, the radiologist has learned how to relate what he sees in radiographs to what he knows about anatomy, together with other information about the patient, to reach a diagnosis. It seems we are a long way from fully understanding these high-level capabilities of human vision. However, they are fundamental to the successful use of the displayed image. The effective coupling to these high-level functions has resulted in some of the biggest achievements in image processing. Computed tomography provides a splendid example. Its success critically hinges on the display of the reconstruction as a proper cross section of human anatomy, which is so easy to interpret that a layman can often see what is wrong. In contrast, the straightforward display of projection data would be impossible for any

human to interpret except for the simplest of objects. See the later example (Fig. 3). The human observer does not seem to be able to efficiently synthesize information over widely spaced regions of an image or, even harder, over different images. The human also cannot effectively deal with complex coded images such as those produced by coded apertures. Although it may not be possible to understand the complex functioning of the human observer in terms of formulae, it behooves the image processor to develop an intuitive understanding of what humans can and cannot see in images. They must develop a sense of the aesthetic in much the same way as artists must. They cannot judge their own results without this understanding. Although some mention is made of the limitations of human vision in the standard textbooks on image processing,[16,36-38] the importance of meeting the needs of the human observer is not emphasized. Obviously, mathematics is more fun. Perhaps the aesthetic aspects of useful images cannot be learned from a textbook, but can only be assimilated through experience, in much the same way as the radiologist needs years of residency to complete his education.

If it were possible to develop a complete mathematical model of human vision, it might be feasible to "optimize" the display of processed images. Unfortunately, such a model still eludes us. Many different kinds of models of human vision have been proposed. Overington[39] has used knowledge about the physiological structure of the eye to correctly predict the contrast sensitivity curve for human observers. It does not seem as though this kind of model can help us design display techniques, because the influence of image noise is not included. Similarly, Cohen, et al.[31] developed an empirical model to explain their contrast sensitivity measurements, including the effects of surround brightness. Baxter et al.[40] have proposed a visual model based in part upon the light adaptation of retinal photoreceptors, which takes into account the surround brightness through occular light scatter. The human contrast sensitivity has been incorporated by Hunt[41] into a constrained least-squares restoration technique. However, its sole effect is to provide another means of regularizing the solution by attenuating the reconstruction at high spatial frequencies, which, it is argued, humans cannot detect anyway. If the results were visually pleasing, it would probably be more because of coincidence than because the contrast sensitivity curve was employed. The effect of random image noise on observer performance has been the subject of the studies by Burgess, et al.[33,34] Human observers were found to be able to perform the given detection tasks almost as well as the ideal observer.[42] This has led to the suggestion[33,34,43] that a variation of the ideal observer may be used as a model for the human. While it seems likely that the human contains some elements of the ideal Bayesian estimator that are operative under ideal display conditions, a number of deficiencies in the human observer must be addressed. These include the inability to detect minute contrast differences and the degradation incurred when the surround brightness is much different than the display brightness. There is room for more observer experimentation because only the simplest detection tasks have been addressed so far. As the specified observer tasks become more complex, the use of information at higher spatial frequencies is required to perform optimally.[44,45] This may hinder the ability of the human to approach the mathematical ideal.

The Displayed Image

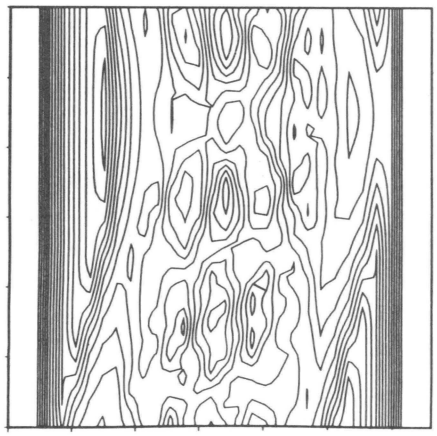
The display of the final image should be fashioned to efficiently couple to the eye-brain system of the observer. It is necessary to overcome the mathematical dictum that it is impossible to completely determine an unknown 2-D function from discrete data. In practice, the display of processed image data is usually made possible by limiting the spatial resolution of the displayed image. This is consistent with the limited visual resolution of the human eye. There is no need to display information at higher spatial frequencies than the observer can see. In spite of the limitation in spatial resolution, for a given set of available data there may still exist a null space and its associated ambiguities. A unique solution may require a further restriction, such as that of minimum norm, as discussed above. There are a number of ways in which the resolution of the display can be limited. Perhaps the simplest is to display the image with a large, but finite number of pixels, sufficient to provide the desired resolution. When the available data have not been sampled with fine enough resolution to allow such an approach, it may be desirable to interpolate between the available samples in order to display a decent-sized image.[46] Such interpolation of a coarsely sampled image does not increase the number of degrees of freedom of the result.[47] It simply offers a more pleasing display of the available data. Alternatively, some reconstruction algorithms, based upon analytic methods, such as filtered backprojection,[48] allow the resolution to be adjusted by selection of the cut-off frequency of a low-pass filter. Other algorithms naturally lead to estimates of the final result that are continuous functions of the spatial variables.[17,49,50] The resolution of the final result can be chosen in many of these also. We will discuss these in conjunction with the examples below.

A number of engineering aspects concerning the physical display system should be considered. Whatever the display system, CRT or film, it is desirable to select the size of the image and the number of pixels to make sure the display meets or exceeds the visual

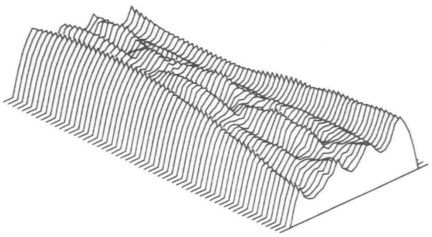
resolution of the observer. For medium-size CRT screens, the number of pixels needed is in the neighborhood of 1000^2 to 2000^2. The display should not flicker, indicating an advantage to noninterlaced over interlaced CTRs and rapid refresh rates (50 or 60 Hertz). Although there seems to be a lack of concern in the medical-imaging community, or even a preference to the contrary, it is reasonable to require that the raster lines abut one another rather than allow interline gaps. This permits the display of a constant function as a constant luminance field, instead of a picket fence. The display should not introduce any graininess or noise itself. This has been a problem in some film/CRT hardcopy systems. The number of gray levels should be sufficient to present a visibly continuous gray-scale. The author feels approximately 256 gray levels is the minimum number required to avoid the contouring effect of too few levels. Also, one should be wary of discontinuous gray-scales that can occur in switching to the next leading binary bit of the video digital-to-analog converter (DAC). In most cases, it seems the use of pseudo-color to display monochrome images can only hinder interpretation. However, color may be useful in the production of eye-catching presentations (for nondiagnostic, but equally important promotional use), the display of quantitative reconstruction values, or the representation of additional degrees of freedom present in the data. In the following examples, the gray-scale images were displayed on a Comtal Vision One/20 with 512^2 pixels and 256 gray levels. The hardcopies were obtained using a LogE/Dunn Instruments, Model 635 camera.

The remaining figures provide examples of some of the important factors in the display of the final image. Figures 3a-f show various ways of displaying the same coarsely sampled projection measurements of the original object, Fig. 3g. Typical computer graphic displays, Fig. 3a and 3b, do not provide the eye with as much information about the raw projection data as a gray-scale display, Fig. 3c, in which one can even "see" evidence of the trajectories of the small circles in the object. However, Fig. 3c is not easy to interpret because compact features in the original object are smeared out along sinusoidal paths. The tomographic reconstructions, Figs. 3d-f, of the original image from these projection data provide an even better visual presentation of the object. Thus, a rearrangement of the available data by means of the reconstruction procedure yields an image that can be interpreted much more readily by the human observer. This demonstrates one of the reasons for the huge success of CT as a medical diagnostic tool. As described in Ref. 50, Fig. 3d is reconstructed using the iterative ART algorithm. The projections of the reconstructed image are obtained by performing the 2-D strip integral over the image under the common assumption that it is composed of piecewise-constant, square regions, called pixels. When the result is displayed in the same way as it is calculated, as square pixels, the blocky appearance is very disconcerting to the eye. A common remedy to this is the use of bilinear interpolation to display the same result, as shown in Fig. 3e. This is somewhat more agreeable to the eye but still possesses visible artifacts arising from the discontinuities in slope inherent in bilinear interpolation. In a more unified approach to reconstruction, it is assumed that the final image is a linear combination of basis functions.[49,50] Such an expansion defines the reconstruction function everywhere. The displayed image is precisely the same as the calculated reconstruction because no interpolation is necessary. When basis functions based upon cubic B-splines are used instead of square pixels, Fig. 3f results. This reconstruction does not possess the undesirable display artifacts of the previous two and provides a reasonable visual indication of which of the four large central objects are squares and which are circles. This is expected to be a difficult discrimination task given the coarse sampling of the projections, as it is known to rely heavily on information at high spatial frequencies.[44,45] Figure 3 demonstrates that the way in which the available data are displayed can greatly influence the amount of information that can be extracted from them. The reason for this has to do with how well each display mode interfaces to the high-level processing of the human brain. CT reconstruction works well because it produces a display with the same morphology as the object, which the eye is accustomed to interpreting.

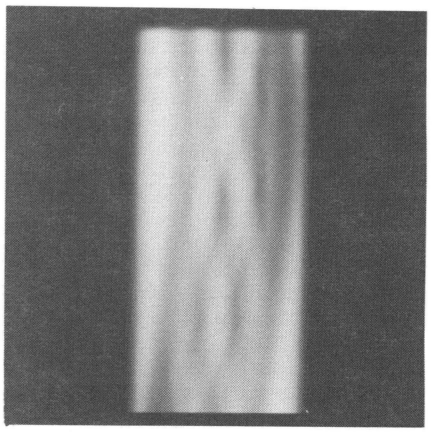
The choice of the spacing and width of the basis functions used to represent the reconstruction directly influence its spatial resolution. Figure 4 shows higher resolution versions of Figure 3e when the measurements consist of line or strip integrals. The improvement in resolution achieved by using a 128 x 128 basis-function grid instead of a 32 x 32 grid permits the result to more closely approximate the measurement-space solution discussed earlier. For the line-integral measurements, the measurement-space solution ideally consists of a linear combination of lines, each with infinitesimal width. Even the approximation to this minimum-norm solution, Fig. 4a, is not visually appealing. The result for strip integrals, Fig. 4b, is better, mostly because the measurements at each angle completely cover the reconstruction region. It would seem to follow that measurements should be designed to achieve full coverage of the region to be reconstructed. It is also concluded that it is best to limit the resolution of the displayed image, as in Fig. 3f, to avoid the appearance of spatial frequencies at which there can be no information in the data, because of the discrete sampling theorem in this case. This is where the strictures of mathematics must be abandoned in favor of producing an aesthetically pleasing image. In limiting the resolution, it is desirable to avoid the



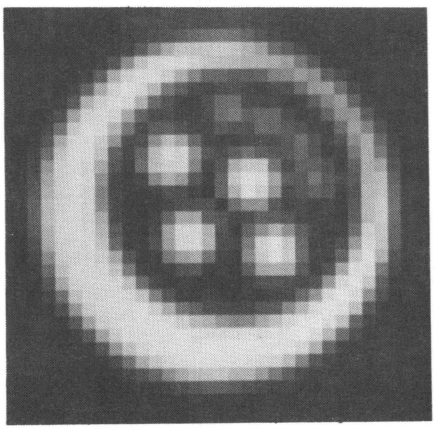
3a



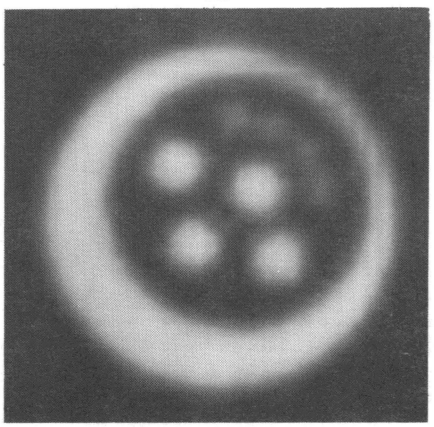
3b



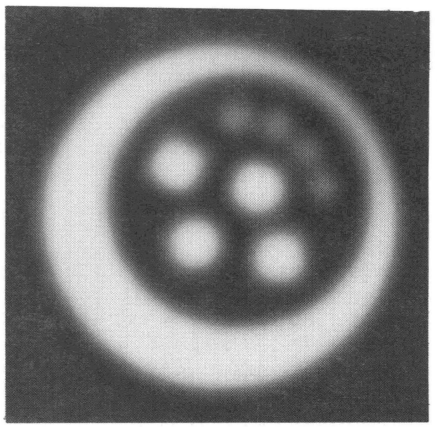
3c



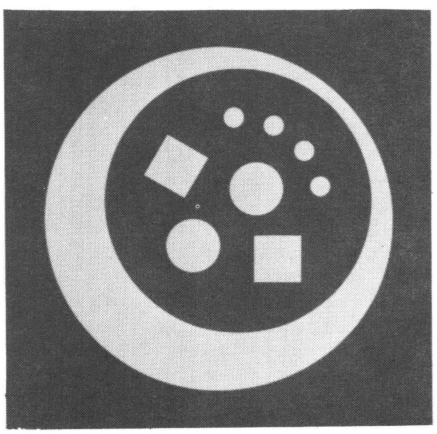
3d



3e



3f



3g

Figure 3. Examples of various methods of displaying coarsely sampled projection data (60 views, 32 samples/view), taken from Ref. 50, in order of improving visual usefulness: a) contour plot of projection data, b) isometric projection (or 3-D relief) of same, c) continuous tone (gray-scale) display of same, d) 32 x 32 pixel reconstruction of original scene from the data using square pixels, e) the same reconstruction displayed using bilinear interpolation, f) 32 x 32 grid reconstruction from the same data using B-spline basis functions, and g) image of original object. The usual graphical displays a) and b) do not show the eye the rich structure contained in the projection data as well as the continuous-tone rendition c) does. Note the disconcerting effects produced by basis functions that either are discontinuous d) or have discontinuous derivatives e). The reconstruction f) comes the closest to providing the eye with the necessary information to ascertain whether the larger, central objects (four sample-spacings wide) are circles or squares.

Figure 4. Reconstructions from same data as in Figure 3 on a 128 x 128 grid using spline basis functions under the assumption that the measurements are a) line integrals and b) strip integrals. The use of the finer grid allows the reconstruction to approach the mathematical pure, measurement-space solution, which may not be visually appealing.

consequences of aliasing[50], namely Moiré effects[16] and ringing at edges, the Gibbs' phenomenon.[36]

Figure 5 shows the improvement in displaying a blurred photograph that can be achieved through image processing. The blur, which was produced by camera motion during the exposure, renders the photographic image very difficult, even impossible to read. Through an enhancement of the recorded information at the appropriate spatial frequencies, the reconstructed image is easily interpreted. This result is obtained using a linear version of the maximum a posteriori probability (MAP) restoration technique[19,51] in which the blurred image is chosen for $\bar{f}$, the ensemble average of f. The similar, well-known Wiener filter would produce a comparable result. Image processing succeeds here because the eye-brain system cannot accomplish the deblurring needed to interpret the information present in the photograph. Incidentally, the recurring artifacts in Fig. 5b arise from the periodic nature of the zeros in the modulation transfer function (MTF) of the blur function and are a result of the null space associated with the blur. Figure 6 presents another example in which a reodering of the available information facilitates better human interpretation. In this case, the radiograph is known to be of an axially symmetric

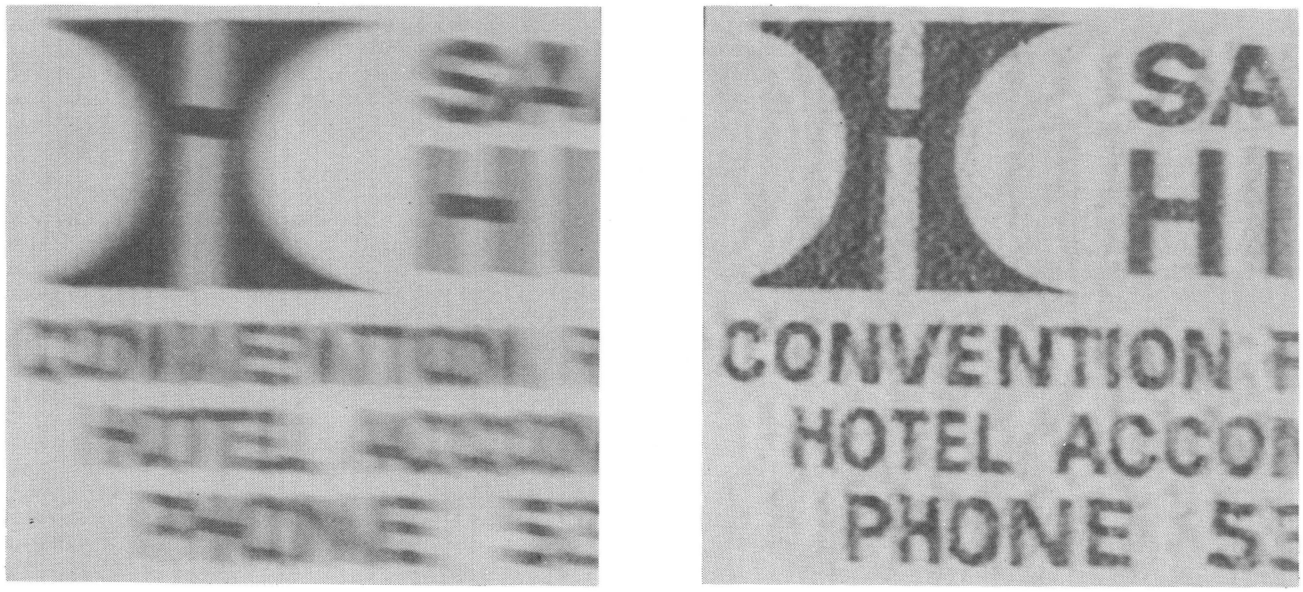


Figure 5. In a typical example of the power of digital image processing, linear MAP restoration of a photograph a) that is subjected to linear-motion blur, produces an eminently readable result b).

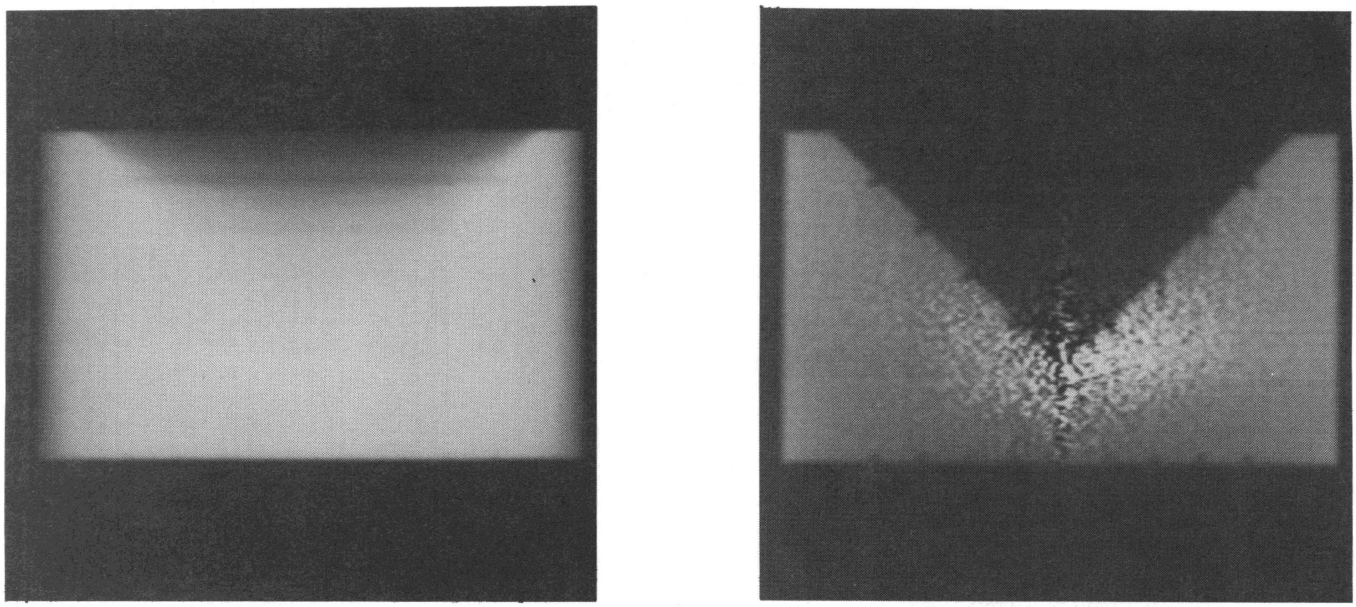


Figure 6. The grooves machined into an axially symmetric object, although not all observable in a) the original radiograph, become visible in b) its tomographic reconstruction (Abel inversion).[52] After reconstruction, interpretation is limited by the noise in the original radiograph, which indicates that the maximum amount of information contained in the radiograph is being extracted.

object. Tomographic reconstruction under the assumption of this symmetry[52] (Abel inversion) yields Fig. 6b. The inner-most grooves in the object cannot be seen in the original radiograph. Their luminosity contrast in directly reading the radiograph is roughly 3.0% on the end face and 1.5% on the inside cone. These grooves are fairly visible in the reconstruction because their contrast is now substantial. Only the noise in the displayed image might hinder their detection. Since this noise originates in the original radiograph, this mode of display is truly approaching the goal of image processing, which is the complete extraction of the information contained in the measurements. These last two examples point out the need to conserve the information inherent in the data through the entire image processing procedure. The early imaging stages should be subjected to particular scrutiny for loss of information. Since the objective of the processing is to overcome the limitations of the observer in viewing the original data, it is unwise to judge the adequacy of the data or of the digitization of input images, as with microdensitometers, on the basis of direct viewing of the unprocessed data.

Image Quality

The term "image quality" is a treacherous one because it seems to mean different things to different people. It is reasonable that image quality should be dependent upon the specified visual task to be performed. What is a good image for one task may not be good for a different task. This diminishes the usefulness of universal figures of merit, which are so often proposed. While the simplicity of these measures of image quality is appealing, they are an oversimplification. In an attempt to define useful and calculable measures of image quality, it has often been assumed that the noise in the image is the limiting factor in interpretation.[42,53-56] Although this may be true of the ideal observer, it is clearly not the case for human observers when the display is inadequate, as illustrated in Figs. 3-6. Nevertheless, such an approach has many redeeming values. It is possible to characterize the information content of an image[55,56] through the straightforward measurement of the optical transfer function[36] (OTF, similar to MTF, but includes scaling of output relative to input) of the imaging system and the noise power spectral density (NPS) in the image. This approach is general because the information content of an image is given as a function of frequency (number of noise-equivalent quanta NEQ(f)). Once the task is specified, it is possible to determine how accurately it can be performed by integrating the NEQ spectrum with the appropriate weighting function.[42,44] The usual and simple engineering definitions of rms noise and signal-to-noise ratio SNR, are woefully inadequate. We learn that SNR is only meaningful when associated with a given task. The SNR may be considered to be a function of spatial frequency. Then the total SNR^2 is just the integral of $SNR^2(f)$ over all frequencies. It follows that typical

filtering operations can only decrease the SNR associated with a specific task because they may discard image information (SNR). Note that because of its unpredictable nature, noise cannot be reduced or eliminated, as is often stated, without affecting the desired signal also. Another advantage of this approach is that the statistical efficiency of transferring information through each step of the imaging chain may be determined. If the statistical efficiency of all the intermediate stages is close to 100%, as it is in CT reconstruction[57], the image quality can be calculated at the measurement stage. This can make the calculation much easier.[58] In this way image quality can be related fundamentally to the initial measurements.

There is obviously a close connection between image quality, defined in terms of task performance, and the model for the human observer. As such a model emerges, it will become clearer how to properly assess image quality. It will be difficult, but necessary, to quantify the more complex, high-level capabilities of the human observer. Only then will we be able to "understand" why the CT reconstruction in Fig. 3f has better image quality for human interpretation than the display of the projection data, Fig. 3c, even though both contain the same information.

Discussion

We have pursued the consequences of the tenet that the aim of image processing is to help the human observer visually interpret image data. Of prime importance is the final displayed image, as that is what the observer looks at. Mathematics plays a fundamental role in guiding image processing, but at times one must transcend mathematics in order to obtain a result. Attention must be paid to the engineering and aesthetic aspects of the display. Without a comprehensive model of the human observer, the selection of the preferred display mode is based more on artistic than scientific grounds. After all the technology, the "eye" is the judge.

In this paper, we have restricted ourselves mainly to image processing. Of course, what has been said about image processing is applicable to imaging as a whole. In fact, it is best to approach any imaging task in a systematic way instead of a piece at a time, as is done so often. Image processing cannot overcome a deficit of information in the measurements. Thus, the requirements of the final image should guide the design of the measurement schema, as it should each step of the imaging chain. An important aspect of a systems approach to diagnosis is the selection of the kind of measurements to take. The radiologist should obviously choose the imaging modality, or combination of modalities,[59] that are most relevant to answering the questions at hand.

Acknowledgments

The author is grateful to Arthur E. Burgess, T. Michael Cannon, Robert F. Wagner, C. Carl Jaffe, George W. Wecksung, and Rollin L. Whitman for many helpful conversations over the years. The help of George Wecksung in developing the CT reconstruction codes that use the basis function approach and in producing the attendant examples shown here is gratefully acknowledged. The original blurred Hilton image, Fig. 5a, was kindly supplied by T. M. Cannon. Darrell A. Terry is responsible for the ray drawing of the CT measurement geometry, Fig. 2. This work was supported by the U. S. Department of Energy under contract number W-7405-ENG-36.

References

1. K. M. Hanson, "Limited angle CT reconstruction using a priori information," Proc. 1st Internat. Symp. Med. Imaging and Image Interp., Berlin, pp. 527-533. 1982.
2. K. M. Hanson, "CT reconstruction from limited projection angles," Proc. SPIE (Appl. Opt. Instr. Med. X), Vol. 347, pp. 166-171. 1982.
3. K. T. Smith, P. L. Solomon, and S. L. Wagner, "Practical and mathematical aspects of the problem of reconstructing objects from radiographs," Bull. Am. Math. Soc., Vol. 83, pp. 1227-1270. 1977.
4. K. M. Hanson and G. W. Wecksung, "Bayesian approach to limited-angle reconstruction in computed tomography," J. Opt. Soc. Am., Vol. 73, pp. 1501-1509. 1983.
5. S. Twomey, "The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurement," J. Frank. Inst., Vol. 279, pp. 95-109. 1965.
6. S. Twomey and H. B. Howell, "Some aspects of the optical estimation of microstructure in fog and cloud," Appl. Opt., Vol. 6, pp. 2125-2131. 1967.
7. S. Twomey, "Information content in remote sensing," Appl. Opt., Vol. 13, pp. 942-945. 1974.
8. B. F. Logan and L. A. Shepp, "Optimal reconstruction of a function from its projections," Duke Math. Jour., Vol. 42, pp. 645-659. 1975.

9. G. T. Herman, and A. Lent, "Iterative reconstruction algorithms," Comput. Biol. Med., Vol. 6, pp. 273-294. 1976.
10. B. F. Logan, "The uncertainty principle in reconstructing functions from projections," Duke Math. J., Vol. 42, pp. 661-706. 1975.
11. C. Hamaker and D. C. Solmon, "The angles between the null spaces of x-rays," J. Math. Anal. Appl., Vol. 62, pp. 1-23. 1978.
12. M. B. Katz, "Questions of uniqueness and resolution in reconstruction from projections," Lecture Notes in Biomathematics, S. Levin Ed., Springer, Berlin, 1979.
13. A. K. Louis, "Ghosts in tomography - the null space of the Radon transform," Math. Meth. Appl. Sci., Vol. 3, pp. 1-10. 1981.
14. B. P. Medoff, W. R. Brody, and A. Macovski, "Image reconstruction from limited data," Proc. Int. Workshop on Physics and Engineering in Medical Imaging, Pacific Grove, Calif., March 15-18, 1982.
15. K. T. Smith, and F. Keinert, "Mathematical foundations of computed tomography," to appear in J. Opt. Soc. Am.
16. A. Rosenfeld, and A. C. Kak, Digital Picture Processing, Academic Press, New York. 1976.
17. H. B. Buonocore, W. R. Brody, and A. Macovski, "Natural pixel decomposition for two-dimensional image reconstruction," IEEE Trans. Biomed. Eng., Vol. BME-28, pp. 69-78. 1981.
18. D. G. McCaughey, and H. C. Andrews, "Degrees of freedom for projection imaging," IEEE Trans. Acous., Speech, Signal Proc., Vol. ASSP-25, pp. 63-73. 1977.
19. B. R. Hunt, "Bayesian methods in nonlinear digital image restoration," IEEE Trans. Comput., Vol. C-26, pp. 219-229. 1977.
20. G. T. Herman and A. Lent, "A computer implementation of a Bayesian analysis of image reconstruction," Inf. Control, Vol. 31, pp. 364-384. 1976.
21. B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," J. Opt. Soc. Am., Vol. 62, pp. 511-518. 1972.
22. A. Lent, "A convergent algorithm for maximum-entropy image reconstruction, with a medical x-ray application," Proc. SPSE Int. Conf. on Image Analysis and Evaluation,. R. Shaw, Ed., pp. 249-257, Toronto, Canada, July 19-23, 1976.
23. G. Minerbo, "MENT: A maximum-entropy algorithm for reconstructing a source from projection data," Comput. Graphics Image Process., Vol. 10, pp. 48-68. 1979.
24. S. J. Wernecke, and L. R. D'Addario, "Maximum-entropy image reconstruction," IEEE Trans. Comp., Vol C-26, pp. 351-364. 1977.
25. S. F. Gull, and G. J. Daniell, "Image reconstruction from incomplete and noisy data," Nature, Vol. 272, pp. 686-690. 1978.
26. R. K. Bryan, and J. Skilling, "Deconvolution by maximum entropy, as illustrated by application to the jet of M87," Mon. Not. R. Astr. Soc., Vol. 191, pp. 69-79. 1980.
27. S. F. Burch, S. F. Gull, and J. Skilling, "Image restoration by a powerful maximum entropy method," Comp. Vis. Graph. Image Process., Vol. 23, pp. 113-128. 1983.
28. J. E. Shore and R. W. Johnson, "Axiomatic derivation of maximum entropy and the principle of minimum cross-entropy," IEEE Trans. Info. Theory, Vol. 26, pp. 26-37. 1980. and Vol. 29, pp. 942-943. 1983.
29. D. M. Titterington, "The maximum entropy method for data analysis," Nature, Vol. 312, pp. 381-382. 1984.
30. S. F. Gull, "Recent developments at Cambridge," to appear in Proc. Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics, Calgory, Canada, August 5-8, 1984.
31. R. W. Cohen, C. R. Carlson, and G. S. Cody, "Image descriptors for displays," Report ONR-CR213-120-2, Office of Naval Research, Arlington, Virginia. 1976.
32. A. J. Alter, G. A. Kargas, S. A. Kargas, J. R. Cameron, and J. C. McDermott, "The influence of ambient and viewbox light upon visual detection of low-contrast targets in a radiograph," Invest. Rad., Vol. 17, pp. 402-406. 1982.
33. A. E. Burgess, and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," J. Opt. Soc. Am., Vol. 1A, pp. 900-905. 1984.
34. A. E. Burgess, and H. Ghandeharian, "Visual signal detection. II. Signal-location identification," J. Opt. Soc. Am., Vol. 1A, pp. 906-910. 1984.
35. D. A. Twible, P. F. Judy, and R. G. Swensson, "Effects of the CT display window on detectability of large and small lesions," Proc. SPIE, Vol. 486. 1984.
36. H. C. Andrews, and B. R. Hunt, Digital Image Restoration, pp. 211-224, Prentice-Hall, Englewood Cliffs, 1977.
37. W. K. Pratt, Digital Image Processing, John Wiley and Sons, New York, 1978.
38. K. R. Castleman, Digital Image Processing, Prentice-Hall, Englewood Cliffs, 1979.
39. I. Overington, Vision and Acquisition, Pentech, London, 1976.
40. B. Baxter, H. Ravindra, and R. A. Norman, "Changes in lesion detectability caused by light adaptation in retinal photoreceptors," Invest. Rad., Vol. 17, pp. 394-401. 1982.
41. B. R. Hunt, "Digital image processing," Proc. IEEE, Vol. 63, pp. 693-708. 1975.
42. R. F. Wagner, "Decision theory and the detail signal-to-noise ratio of Otto Schade," Photogr. Sci. Eng., Vol. 22, pp. 41-46. 1978.

43. R. F. Wagner and D. G. Brown, "More unified analysis of medical imaging system SBR characteristics," Proc. SPIE (Appl. Opt. Instr. Med. XII), Vol. 454, pp. 2-8. 1984.
44. K. M. Hanson, "Variations in task and the ideal observer," Proc. SPIE (Appl. Opt. Instr. Med. XI), Vol. 419, pp. 60-67. 1983.
45. K. M. Hanson, "Optimal object and edge localization in the presence of correlated noise," Proc. SPIE (Appl. Opt. Instr. Med. XI), Vol. 454, pp. 9-17. 1984.
46. H. S. Hou and H. C. Andrews, "Cubic splines for image interpolation and digital filtering," IEEE Trans. Acous. Speech Sig. Proc., Vol. ASSP-26, pp. 508-517. 1978.
47. H. C. Andrews and C. L. Patterson, III, "Digital interpolation of discrete images," IEEE Trans. Comp., Vol. C-25, pp. 196-202. 1976.
48. L. A. Shepp and B. F. Logan, The Fourier reconstruction of a head section," IEEE Trans. Nucl. Sci., Vol. NS-21, pp. 21-43. 1974.
49. H. S. Hou and H. C. Andrews, "Least squares image restoration using spline basis functions," IEEE Trans. Comp., Vol. C-26, pp. 856-873. 1977.
50. K. M. Hanson, and G. W. Wecksung, "Local basis-function approach to computed tomography," submitted to J. Opt. Soc. Am.
51. T. M. Cannon, H. J. Trussell, and B. R. Hunt, "Comparison of image restoration methods," Appl. Opt., Vol. 17, pp. 3384-3390. 1978.
52. K. M. Hanson, "Tomographic reconstruction of axially symmetric objects from a single radiograph," Proc. 16th Inter. Congress on High Speed Photography and Photonics, Proc. SPIE, Vol. 491, Strasbourg, August 27-31, 1984.
53. R. Shaw, "Photon fluctuations, equivalent quantum efficiency and the information capacity of photographic images," J. Phot. Sci., Vol. 11, pp. 313-320. 1963.
54. H. J. Zweig, "Performance criteria for photodetectors - concepts in evolution," Phot. Sci. Eng., Vol. 8, pp. 305-311. 1964.
55. K. M. Hanson, "Detectability in computed tomographic images," Med. Phys., Vol. 6, pp. 441-451. 1979.
56. R. F. Wagner, D. G. Brown, and M. S. Pastel, "Application of information theory to the assessment of computed tomography," Med. Phys., Vol. 6, pp. 83-94. 1979.
57. K. M. Hanson, "On the optimality of the filtered backprojection algorithm," J. Comput. Assist. Tomo., Vol. 4, pp. 361-363. 1980.
58. R. F. Wagner, D. G. Brown, and C. E. Metz, "On the multiplex advantage of coded source/aperture photon imaging," Proc. SPIE (Conf. Digital Radiog.), Vol. 314, pp. 72-76. 1981.
59. C. C. Jaffe, "Editorial-Integrated medical imaging," Invest. Rad., Vol.14, pp. 1-3. 1979.