A Survey of Coded Aperture Imaging

BY

H. H. Barrett
Optical Sciences Center
University of Arizona
Tucson, Arizona

I. Introduction

There are a number of important applications, particularly in nuclear medicine and astronomy, where it is necessary to form an image of a source emitting x-rays or gamma rays. Since this radiation is not appreciably refracted or reflected by matter, the image cannot be produced by ordinary focusing optics. Instead, the conventional approach has been to use a pinhole aperture or multihole collimator as shown in Fig. 1 to form an image of the source on a scintillation camera, film, or other image detector.

This type of imaging has made possible the field of diagnostic nuclear medicine. With a modern scintillation camera, acceptable images with 1 to 2 cm resolution can be obtained with a radiation dose to the patient that is one to three orders of magnitude below the dangerous level. The information gained in this manner is valuable in delineating the physiological function of an organ, e.g., which parts of the organ absorb a given pharmaceutical and how rapidly. However, the resolution is inadequate to give the kind of detailed, morphological information contained in a radiograph. Improved resolution in nuclear medicine would offer many new diagnostic possibilities, such as the early detection of small focal lesions and metastases, or more precise differentiation between benign and malignant tumors.

But increasing resolution with a pinhole or collimator is a very difficult task. Because of the low collection efficiency of the apertures (typically 0.01%), the scintillation camera must have high detection efficiency in order to "make every photon count." The high efficiency is obtained by using a relatively thick scintillation crystal and sensitive photomultipliers to detect the resulting light. Economic constraints limit the number of photomultipliers and hence the intrinsic spatial resolution of the camera. But even if that problem were solved, the resolution would still be severely limited by the collimator itself. For example, suppose it is desired to increase the system resolution from 12 mm to 6 mm and that a fictitious camera with perfect intrinsic resolution is available. Then the resolution can be increased a factor of two simply by using a finer collimator or smaller pinhole. With all other factors remaining constant, the result is a four-fold reduction in gamma ray flux. However, in order to keep the same signal-to-noise ratio, the number of detected gamma ray quanta per resolution element must remain constant, or the total number of quanta should increase by a factor of four. The net result is that a factor of two increase in resolution must be accompanied by a factor of sixteen increase in patient dose or exposure time.

A similar result is obtained for a scintillation scanner used with a focused collimator. The scanner has an efficiency of 1 to 5%, but it views only one resolution element of the object at a time. Again large doses or exposure times, which can be reduced only at the expense of spatial resolution, are required. For example, a typical whole-body skeletal scan now takes about an hour.

Thus, whether a camera or scanner is used, an apparently inevitable tradeoff between radiation flux and resolution is required. Very similar tradeoff situations occur also in other fields such as radar technology,[1] x-ray astronomy,[2-4] and infrared spectroscopy.[5-8] But in these areas sophisticated signal processing techniques have been devised to avoid the tradeoff, to obtain high flux, and after the processing, high resolution as well.

These ideas have recently been carried over into nuclear medicine,[9-11] where they have come to be called "coded aperture imaging" techniques. The key feature that sets coded aperture imaging apart from conventional imaging techniques is that the initial image is not immediately recognizable. Instead it is a coded or scrambled representation of the object. The processing step is not merely to enhance the image, but is essential to obtain any useful image at all.

The most successful coded aperture to date is the Fresnel zone plate.[2, 3, 9-11] In this case the coded image is closely analogous to a hologram and can be reconstructed with a coherent optical data processing system.

The zone plate aperture can be a factor of 1000 more efficient than a pinhole or parallel-hole collimator. For a point source, this means that an image with a given signal-to-noise ratio can, in theory, be formed with a factor of 1000 smaller exposure time or dose.[12, 13] For larger sources, the advantage is considerably less, but it should still be as much as a factor of 10 for many practical clinical situations.[13]

However, in most of the work we have performed to date, this large advantage is not realized, primarily because x-ray film, rather than a scintillation camera, is used as the detector.[11, 14] The resulting camera is an extremely simple one that is lightweight, portable and inexpensive, and gives high-resolution, tomographic[15] images. Clinical trials of this camera are now under way.[16] Excellent results have been obtained in studies of the thyroid, bone lesions, myocardium, lungs, liver, and spleen. Examples are shown in Figs. 2 and 3.

Although this camera definitely seems to be a useful diagnostic device, it requires substantially larger doses or exposure times than conventional systems. Until recently, it had seemed that it might not be possible to realize the full theoretical advantage of coded apertures because an off-axis zone plate[11, 14, 15] seemed to be necessary to eliminate undiffracted background light. The off-axis zone plate has very closely spaced zones and is not, therefore, compatible with a low-resolution, high sensitivity, detector such as the Anger camera. Thus an impasse had developed: To utilize the high sensitivity of the zone plate it was necessary to use a low-sensitivity detector like x-ray film, so the overall sensitivity was no better, and usually substantially worse, than with conventional systems.

Recent work, however, has shown that this impasse may be surmounted in several ways. As discussed below, several apertures including the on-axis zone plate, annulus and non-redundant pinhole arrays can be used with an Anger camera.

II. Basic Principles of Zone Plate Imaging

The operation of the zone plate aperture can best be understood by first considering a point source of gamma rays. The probability distribution of the gamma rays on the image plane is just a geometric shadow of the zone plate (see Fig. 4). This shadow contains the full three-dimensional information about the location of the point; the lateral position of the shadow depends linearly on the lateral position of the point, and the scale of the shadow depends on the distance of the point from the zone plate.

Thus the imaging operation is a transformation of each point in the object into a zone plate. Exactly this transformation occurs in optical holography where the zone plate is the interference pattern between the light scattered from an object point and the reference beam. Therefore, once the coded image has been recorded, it no longer matters how the zone plates were formed. Whether it was by geometric shadowing of gamma rays or interference of light waves, a hologram is obtained.

The reconstruction step uses the fact that a zone plate transparency behaves as a lens, but with diffraction rather than refraction deviating the light. Thus, if the zone plate shadow is recorded as a transparency of suitable scale and inserted into a laser beam, part of the light is diffracted into a bright local spot that is the reconstructed image of the gamma ray point source. If two adjacent points sources were present (Fig. 4), the hologram would consist of overlapping zone plate shadows, but the reconstruction would be two non-overlapping focal spots. Two sources at different distances from the zone plate would produce foci in two different planes.

Then, to the extent that linear superposition holds, any extended object can be considered to be composed of points and should be properly imaged. However, in practice this seldom occurs except for small objects. The difficulty is the undiffracted light (or "dc light" in optical jargon) that is present since both the gamma ray source intensity and the zone plate transparency are non-negative functions. For a point source, this undiffracted light overlaps the focal spot, but it is much less intense and therefore not troublesome. When a second point source is added, the two focal spots may be well resolved, but the undiffracted light patterns are not. Therefore the presence of a second point reduces the contrast of the first. For large, extended sources, the image contrast becomes so small that the image can no longer be seen.

A spatial filtering technique to alleviate this problem has been described.[9] In this approach, a Schlieren stop is used to block low-spatial frequency components of the image, thereby suppressing the undiffracted light. However, the usefulness of this technique is limited because a large object itself has low frequency components. Therefore, a stop large enough to block the undiffracted light will also block much of the desired object information.

A better approach is to use an off-axis section of a zone plate with a halftone screen plate, the undiffracted light pattern does not overlap the local spot and is therefore not a problem. The halftone screen is necessary because the off-axis zone plate will not properly code low spatial frequencies. The screen serves to spatially heterodyne the object spatial frequency spectrum into the passband of the zone plate. When the zone plate and halftone frequencies are properly matched there is a strong Moire fringe pattern formed on the image plane as shown in Fig. 5. Diffraction from these fringes forms the reconstruction.

The only major drawback to the off-axis zone plate is that its rings are very closely spaced. Therefore, a high-resolution image detector must be used. Alternatively, for a given detector resolution, the resolution of the final image will be much worse with an off-axis plate than an on-axis one. As a useful rule of thumb,[12] the final image resolution with an off-axis zone plate will be about a factor of three worse than the intrinsic detector resolution. Therefore, an Anger camera is not very attractive, and we turned to x-ray film as the detector.

An experimental camera consisting simply of an off-axis zone plate, a halftone screen, and an x-ray film cassette was built several years ago. The zone plate and halftone were prepared either by casting lead onto tinned aluminum plates and machining them into the desired pattern, or by electroforming gold.

The output of the camera is a low-density coded picture on x-ray film, usually 10 x 12 in. This film is placed in a copy camera and copied at reduced scale onto a photographic plate, which is then developed and bleached so that the original density differences are translated into optical phase differences.

The reconstruction or decoding system, shown in Fig. 6, uses a helium-neon laser, a lens-pinhole spatial filter, a collimating lens, and an iris to select the correct diffraction order of the hologram. The reconstructed image first appears at a very small scale and is magnified by a subsidiary lens onto a fine ground glass. The ground glass may be viewed visually, or with a silicon vidicon TV camera.

An important part of the reconstruction system is the iris that separates the correct diffraction order from the undiffracted light and the other orders. The open diameter of the iris determines the bandwidth of the system. Image smoothing is therefore easily accomplished if desired.

The point at which the reconstruction occurs along the axis of the optical system depends on the focal length of the reduced coded image, which in turn depends on the distance of the object from the zone plate. Thus, the image may be examined plane by plane by moving the magnifying lens in the reconstruction system. All planes of the object are thus contained in a single coded image.

Examples of the clinical images obtained with this system are shown in Figs. 2 and 3.

III. Background Suppression Techniques

The main problem in coded-aperture imaging is that the apertures always have a positive-definite transmission. In zone plate imaging with optical processing this leads to the undiffracted or "dc" light background in the reconstructed image. For point objects, this undiffracted component is relatively unimportant since it is weak compared to the desired diffracted light. However, for large continuous objects it completely obscures the reconstructed image.

Two solutions to this problem have already been mentioned. Some improvement can be obtained by spatial filtering[9] with a dc stop in the Fourier plane, but, in general, high quality imagery cannot be obtained this way. A stop large enough to block all of the dc term will frequently also block most of the signal. Another solution is to use an off-axis section of a zone plate, for which the dc and signal components are spatially separated. Here it is necessary to also use a halftone screen as a spatial carrier. The drawback to this approach is that the off-axis zone plate has very fine rings, placing stringent requirements of the spatial resolution of the image detector.

Still another approach, suggested by Tipton,[30] is to choose the zone plate in such a way that the phase of the light in the reconstructed image is opposite to that of the background. The result if a dark image on a bright field. The drawbacks to this method are that it works only for a restricted range of object sizes and that the inevitable nonuniformities in the background can be misinterpreted as nonuniformities in the object.

It has long been recognized[18-20] that an effective dc suppression could be obtained if two coded images were formed, one with a positive zone plate and one with its negative, and these coded images were subtracted before reconstruction. Let the transmission of the positive zone plate, treated as sinusoidal, be denoted by

$$g_+(\vec{r}) = \frac{1}{2} + \frac{1}{2} \sin ar^2, \quad (r \le R)$$

and that of the negative one be

$$g_-(\vec{r}) = \frac{1}{2} - \frac{1}{2} \sin ar^2, \quad (r \le R)$$

where $\vec{r}$ is a two-dimensional position vector, $r = |\vec{r}|$,

$$\alpha = \pi / r_1^2,$$

and r_1 is the radius of the first zone.

The coded image is basically a convolution[12] of the aperture transmission with the object distribution $f(\vec{r})$. Let the individual coded images, obtained with the positive and negative zone plates, be given by $h_+(\vec{r})$ and $h_-(\vec{r})$, respectively, where

$$h_\pm(\vec{r}) = f(\vec{r}) \ast g_\pm(\vec{r})$$

The asterisk denotes a convolution with suitable scale factors as discussed in Ref. 12. The difference image, $h(\vec{r})$, is

$$h(\vec{r}) \equiv h_+(\vec{r}) - h_-(\vec{r})$$

$$= \frac{1}{2} f(\vec{r}) \ast \sin ar^2$$

Thus the net effect is a convolution with a bipolar (positive and negative valued) zone plate, even though negative transmissions have no physical meaning. A bipolar zone plate has no constant or dc term in its transmission and therefore does not give rise to a dc background term in the reconstruction.

This approach has recently been implemented with computer processing by MacDonald et. al.[18] and with optical processing by Stoner, Wilson and Barrett.[21, 22] In the latter case, the subtraction was performed by means of a grid coding technique originally proposed by Pennington et. al.[23]

The basic arrangement for grid-coded subtraction with positive and negative zone plates uses a standard Anger camera without any modification except for a grid (Ronchi ruling) which is placed adjacent to the film in the oscilloscope camera. A coded image is first recorded with a positive zone plate in place. Then the negative zone plate is substituted, the grid is shifted by half its period and a second coded image is recorded. Since the film in the oscilloscope camera is not moved between exposures, the two coded images are interleaved on the film. A reduced-scale copy of this film is inserted into the optical system shown in Fig. 6. When a single diffraction order is selected with the iris, it can easily be shown that the dc components cancel while the signal components add.

The first attempt to experimentally demonstrate this technique with an Anger camera gave very encouraging results. Subsequent experiments, however, soon showed that one problem remained. Although the dc component was effectively cancelled, there was still an out-of-focus twin image term that substantially degraded the image for large objects. The solution to this problem was found in the literature on optical holography. Adapting an idea of Burckhardt and Doherty,[24] we went to a succession of three or four zone plates (e.g., positive and negative sine and cosine plates). Then it could be shown[22] that both dc and twin image terms cancelled and that the result was mathematically equivalent to using an off-axis zone plate and halftone screen. This result was fully verified by optical simulation experiments; Anger camera experiments are in progress.

It rapidly became apparent that the grid-coded subtraction technique opened up a great variety of new possibilities for coded apertures. To date, we have demonstrated, by optical simulation, the feasibility of using an annular aperture, an inverted zone plate, a spiral zone plate, a step-wise spiral zone plate, and the Girard grill.[22]

IV. Quantum Noise in Coded Aperture Systems

Although a coded-aperture can have up to a thousand times the collection efficiency of a pinhole or collimator, it has been recognized from the beginning that this did not imply that dose or exposure time could be reduced by such a factor. The difficulty is that although coded aperture techniques collect more quanta, they also require more quanta to get the same signal-to-noise ratio.

Very detailed analyses of quantum noise have been presented, both for coded aperture imaging per se[12, 23, 25-27] and for the analogous area of multiplex spectroscopy.[28, 29] A full account of the mathematical details is beyond the scope of this review, but the basic results can be stated fairly simply. Consider an assembly of m point sources of equal strength and assume that each point is well resolved from all the others, but that the whole assembly is sufficiently compact[50] that all of the individual zone plate shadows almost completely overlap. Then it is fairly easy to show[12, 13] that the signal-to-noise ratio (SNR) for the quantum-limited zone plate camera is given by

$$(\mathrm{SNR})_{zp} = \frac{\alpha\sqrt{N_t}}{m}$$

where N_t is the total number of detected quanta and α is a factor somewhat less than unity whose precise value depends on the details of the data processing. For comparison, the SNR for an equivalent pinhole is given by

$$(\mathrm{SNR})_{ph} = \sqrt{N_t/m}.$$

Since N_t increases linearly with exposure time T in both cases, it is easy to see that the ratio of exposure times for constant SNR in the two cases is

$$\frac{T_{ph}}{T_{zp}} = \frac{\alpha^2}{m}\,\frac{(N_t)_{zp}}{(N_t)_{ph}} = \frac{\alpha^2}{m}\,\frac{\eta_{zp}}{\eta_{ph}}$$

where η denotes geometric efficiency. This equation shows that for a single point source (m = 1) almost the full efficiency advantage of the zone plate can be translated into an exposure time or dose advantage. As the object becomes larger and m increases, the advantage progressively decreases. For very large objects, such that the individual shadows no longer completely overlap, Eq. (3) must be modified, but the exact theory[13] still shows that there can be a net disadvantage to the zone plate for large objects. Similarly, the zone plate can be a disadvantage in imaging weak sources adjacent to strong ones. In general terms, the zone plate should have an advantage when imaging areas of the object that are substantially stronger than the average, taken over the field of view of the camera. Thus, thyroid, kidneys, brain, bone, and myocardium should all be "good" objects for zone plate imaging; lungs, liver, and placenta are "bad."

These quantum noise considerations constitute an inherent limitation to coded aperture techniques. If it were the only limitation, coded apertures could still be used to reduce patient dose or exposure time in most nuclear medicine imaging procedures. However, in practice, other factors creep in to reduce the sensitivity drastically, especially if film is used as the detector. These factors include film grain noise, scattered radiation,[16] incomplete absorption of the gamma rays, the absorption of the half-tone screen[25] and inefficient data processing.[13] However by using an Anger camera as the detector, computer processing and any of several apertures that do not require a half-tone screen, it is now possible to obtain the full theoretical advantage of coded apertures.

V. Current Research

Three trends can be discerned in current research in coded-aperture imaging. These are: (1) Use of time-varying apertures; (2) Use of "dilute" apertures with transmission much less than 50%; and (3) Attempts to derive transverse tomographic sections, unblurred by other planes, from coded images. Each of these pursuits has met with some degree of success, as we will now discuss.

We have already seen an example of temporally modulated coded apertures in the multiple zone plate system described in sec. III. In addition, Macovsky[30] has proposed a pinhole array in which each pinhole is opened and closed with a different temporal frequency. Spectral analysis serves to unravel the resulting image.

Workers at the University of Michigan[31] have used time-modulated "stochastic" apertures. Basically this approach employs a one-dimensional code whose cross-correlation with a cyclically repeated replica of itself has a uniform background level. A two-dimensional pinhole array is then temporally modulated by this code, so that each pinhole executes the code with a different phase. Digital decoding is used. The main advantage of this approach seems to be that out-of-focus planes blur smoothly without generating the coherent artifacts that can be troublesome with zone plates.

Two examples of dilute apertures have been discussed in the literature--the annular aperture and the non-redundant[32] pinhole array. The annulus was first proposed by Walton[33] who also suggested a clever video processing scheme. Mathematically, his processing method amounts to correlating the coded image with a very thin annulus of appropriate scale. This works well for point objects but fails for larger objects[34] because the point spread function falls off only as $1/r$, where r is the radial distance from the true location of the point. The reconstructed intensity at any point is therefore dependent on the object intensity at very distant points. Recent work by Simpson et. al.[34] shows that this problem may be avoided by suitable filtering.

The non-redundant pinhole apertures[32] are arrays in which the vector distance between any two pinholes is not repeated. They have superior autocorrelation properties to the completely random pinhole arrays proposed by Dicke.[4] Both digital and optical decoding has been employed.

Dilute apertures, in general, are attractive because they do not impose stringent requirements on the count rate capability, dynamic range or spatial resolution of the detector. They collect fewer counts than a filled aperture such as a Fresnel zone plate, but also require less for the same signal-to-noise ratio.[34]

Attempts to derive transverse section data from coded images have been inspired by the recent success for the EMI-scanner. Work in this area is underway at the Universities of Michigan and Arizona and the Lawrence Berkeley Laboratory.[35] The latter group in particular, has demonstrated that an image of a single isolated plane can be obtained with a pinhole array looking down on the plane. There is no significant degradation due to other planes.

REFERENCES

1. J. R. Klauder, A. C. Price, S. Darlington, and W. J. Albersheim, Bell System Tech. J. 39, 745 (1960).
2. L. Mertz and N. O. Young in Proc. Internat. Conf. on Optical Instruments, (Chapman and Hall, London, 1961) p. 305.
3. N. O. Young, Sky and Telescope 25, 8 (1963).
4. R. H. Dicke, Astrophys. J. 153, L101 (1968).
5. L. Mertz, Transformations in Optics (Wiley, New York, 1965) p. 88.
6. A. Girard, Appl. Opt. 2, 79 (1963).
7. M. J. E. Golay, J. Opt. Soc. Am. 39, 437 (1949); 41, 468 (1951).
8. M. Harwit, Appl. Opt. 10, 1415 (1971).
9. H. H. Barrett, J. Nucl. Med. 13, 382 (1972).
10. W. L. Rogers, K. S. Han, L. W. Jones, and W. H. Beierwaltes, J. Nucl. Med. 13, 612 (1972).
11. H. H. Barrett, G. D. DeMeester, D. T. Wilson, and M. H. Farmelant in Medical Radioisotope Scintigraphy 1972 (Vienna, IAEA, 1973) Vol. I, pp. 269-284.
12. H. H. Barrett and F. A. Horrigan, Appl. Opt. 12, 2686-2702 (1973).
13. H. H. Barrett and G. D. DeMeester, Appl. Opt. 13, 1100 (1974).
14. H. H. Barrett, D. T. Wilson, G. D. DeMeester, and H. Scharfman, Opt. Eng. 12, 8 (1973).
15. H. H. Barrett, G. D. DeMeester, D. T. Wilson, and M. H. Farmelant in Tomographic Imaging in Nuclear Medicine, G. S. Freedman, Ed. (New York, Soc. Nucl. Med., 1973) p. 120.
16. M. H. Farmelant, G. D. DeMeester, D. T. Wilson, and H. H. Barrett, J. Nucl. Med., to be published.
17. M. Tipton, F. Bonte, and H. J. Caulfield, Opt. Eng. 12, (1973).

18. B. MacDonald, IEEE Nuclear Science Symposium, San Francisco (1974).

19. L. Mertz, Proc. IRIS 5, 157 (1960).

20. L. Mertz, Advertisement, J. Opt. Soc. Am. Feb. 1960.

21. W. W. Stoner, D. T. Wilson, and H. H. Barrett, Opt. Comm., to be published.

22. H. H. Barrett, W. W. Stoner, D. T. Wilson, and G. D. DeMeester, SPIE Annual Meeting, San Diego (1974) and Opt. Eng. 13, 539 (1974).

23. K. S. Pennington et. al, Opt. Comm. 2, 113 (1970).

24. C. B. Burckhardt and E. T. Doherty, Appl. Opt. 7, 1191 (1968).

25. A. Lindgren et. al., International Symposium on Optical Computing (Zurich, 1974).

26. G. Levy, J. Nuc. Med. 15, 214 (1975).

27. H. H. Barrett, J. Nuc. Med. 15, 215 (1974).

28. C. J. Oliver and E. R. Pike, Appl. Opt. 13, 158 (1974).

29. V. I. Tsoi and T. N. Sokolova, Opt. Spec. 36, 240 (1974).

30. A. Macovsky, Phys. Med. Biol., to be published.

31. R. May, et. al., Appl. Opt. 13, 2589 (1974).

32. L. T. Chang et. al., to be published.

33. P. Walton, J. Nucl. Med. 14, 861 (1973).

34. R. G. Simpson, H. H. Barrett and H. D. Fisher, Int. Opt. Computing Conf. (Washington D. C.).

35. B. MacDonald et. al., Int. Opt. Computing Conf. (Washington D. C.), 1975.

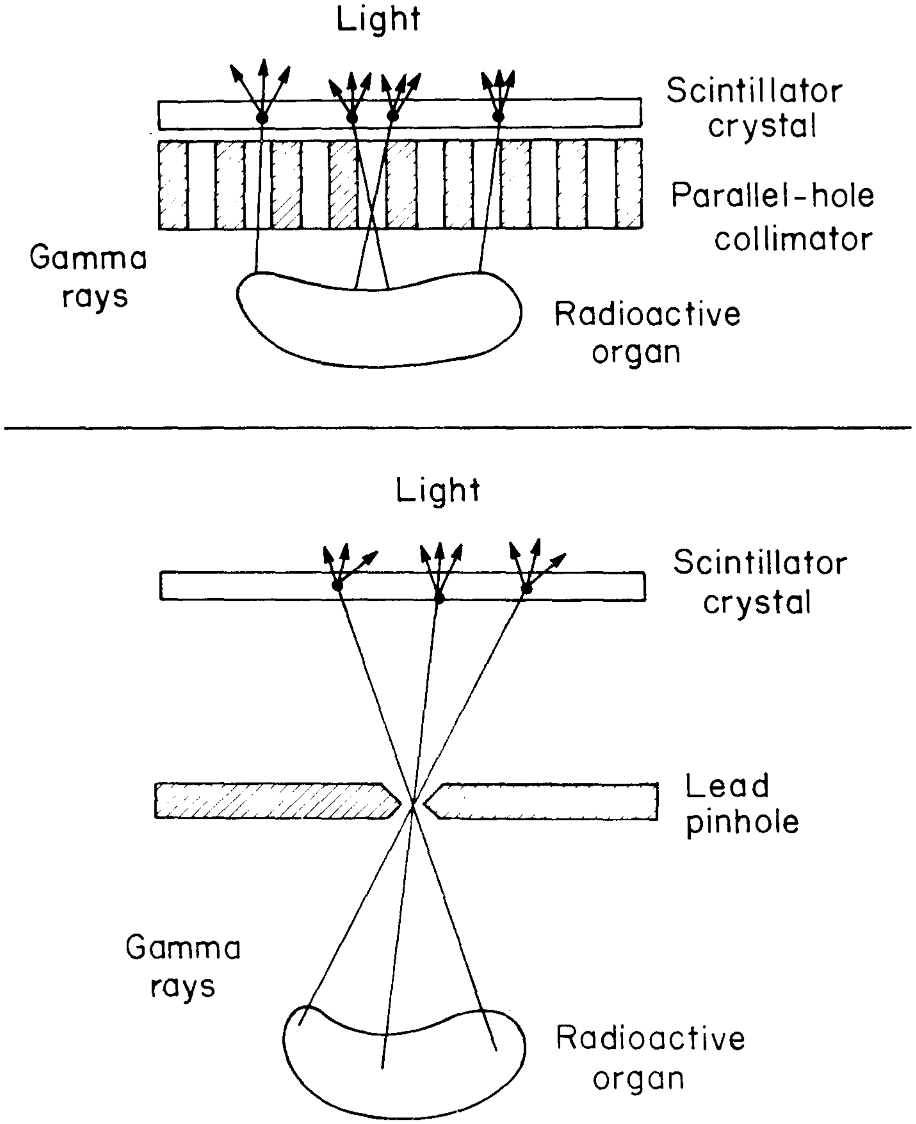

Figure 1. Conventional methods of gamma-ray image formation in Nuclear Medicine.

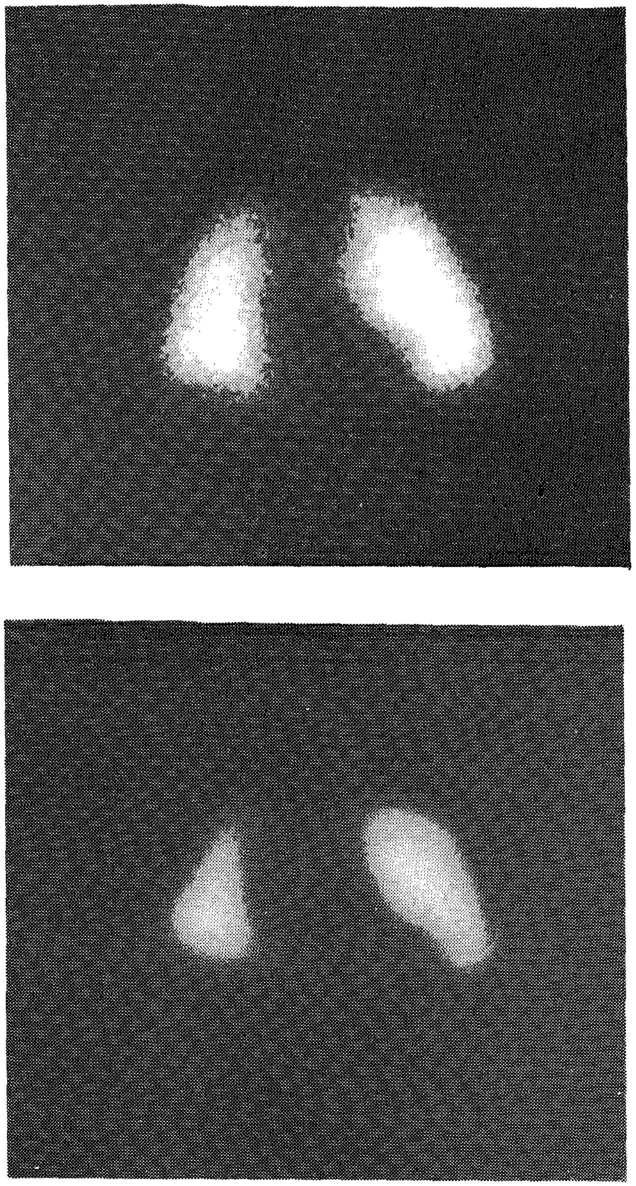

Figure 2. Normal lung perfusion images obtained with a conventional scintillation camera and collimator (left) and with the zone plate/film cassette camera (right).

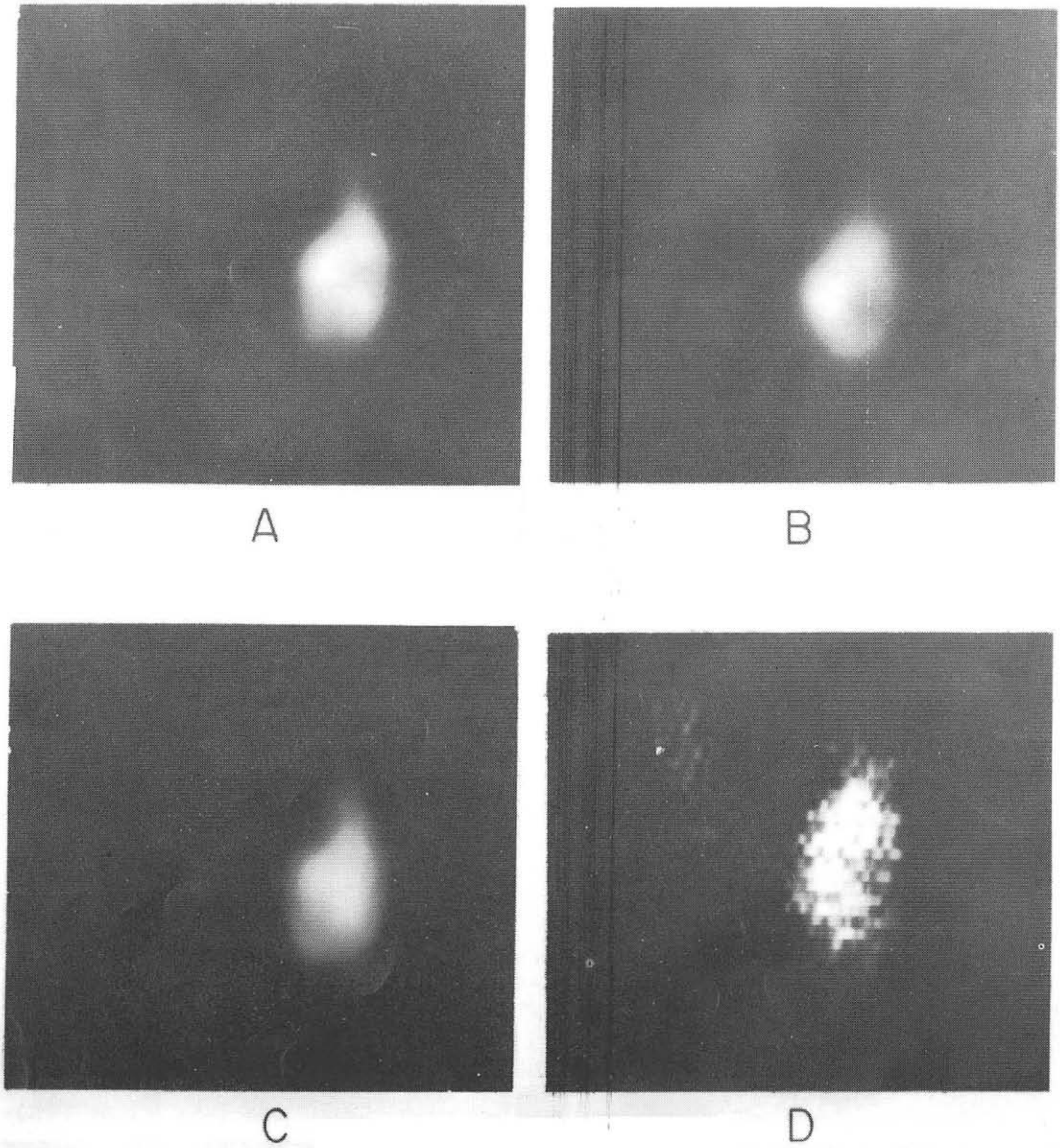

Figure 3. Images of the thyroid gland of a patient who had undergone a partial thyroidectomy. A-C are three different tomographic levels, separated by 6 mm, as imaged with the zone plate camera. D was taken with a rectilinear scanner.

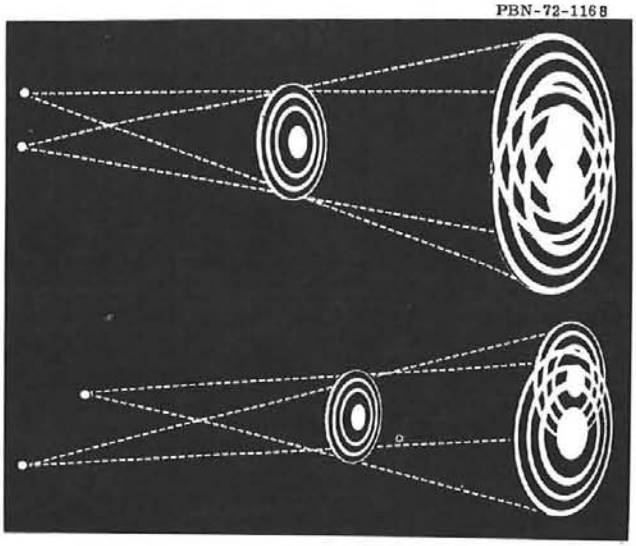

Figure 4. Basic principle of the Mertz and Young reticle camera.

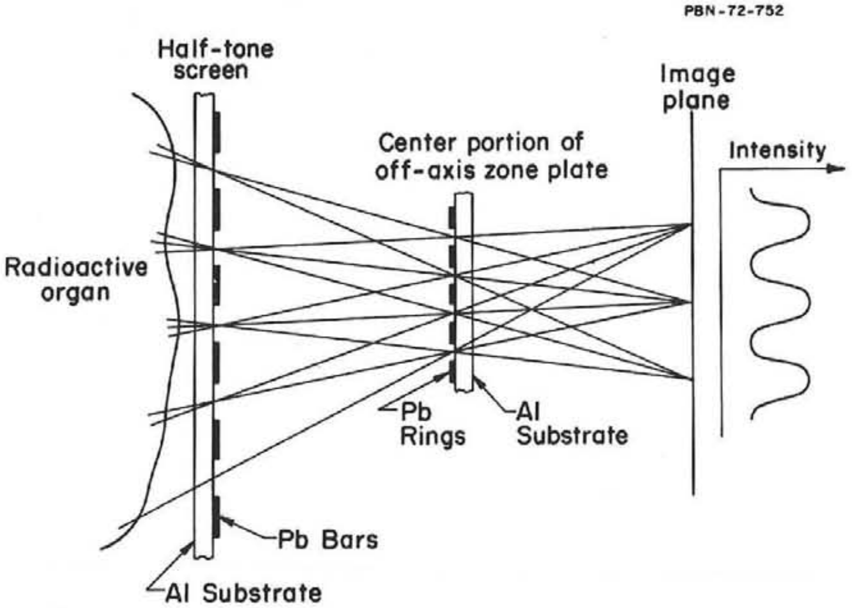

Figure 5. Illustration of the formation of Moire fringes by a zone plate and half-tone screen.

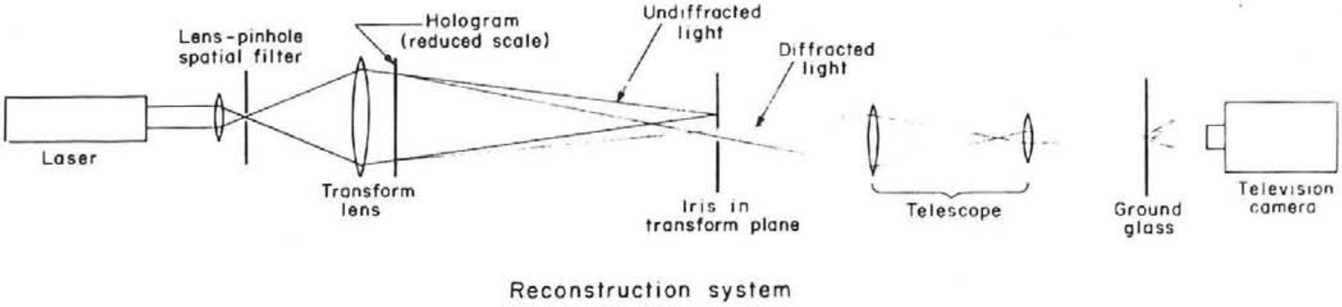

Figure 6. Optical reconstruction system