Visual interpretation of complex data

by E. J. Farrell

With increasingly complex digital simulations and computations, larger volumes of output are generated, and users must select a concise method of displaying the output and extracting relevant information. A set of imaging functions and display modes is developed to interpret data effectively for a wide range of applications. The imaging functions are complementary. Each function is useful for a different aspect of data interpretation. The relationships between variables and the global structures within the data are obtained with different display modes such as multiple windows and animation. With this set of complementary imaging functions and display modes, more information is obtained than with prior imaging methods. Also, more complex simulation studies are feasible since the results can now be visualized.

As digital simulations and computations become more complex, larger volumes of output are generated, often several million bytes of multidimensional data for each study. System users are faced with the difficult problem of selecting a concise method of displaying the output and extracting relevant information. Simple display methods based on contour plots, three-dimensional wire-frame images, or black-and-white halftone images do not have sufficient capacity. New methods are required to interactively display and interpret three- and four-dimensional data.

The value of multidimensional color graphics and imaging for interpreting complex data has been demonstrated by several authors. Graedel and McGill use four methods to display results from a chemical kinetics computation: contour plots, two-dimensional (2D) chromatic plots, surface plots of 2D data, and three-dimensional (3D) binary plots.[1] They illustrate the value of these methods for interpreting 2D and 3D data on atmospheric air quality. Brown presents several 2D and 3D graphical methods for finite-element mechanical design.[2] Graphics are used for defining the computation model (node mesh) as well as for interpreting the results. Computed parameters (strain, temperature, pressure, displacement, etc.) are displayed in relation to the computation mesh. Farrell, Laux, Corson, and Buturla demonstrate how 3D imaging, animation, and multiple display windows facilitate interpretation of multivariable data.[3] Two- and three-dimensional simulations of solid state devices are used to illustrate the imaging methods, which include time-varying 3D arrays and vector flux data. Mohr and Vaughan use several graphical methods to display multivariable thunderstorm measurements.[4] Their data require interpolation of irregularly spaced scalar and vector data to a rectilinear grid prior to display and interpretation.

These prior studies focus on methods of displaying data for particular applications. In the research presented here, a set of imaging functions and display modes is developed to interpret data effectively for a wide range of applications. The imaging functions are complementary; each function is useful for a different aspect of data interpretation. For example, the user may wish to rapidly review all of the data using gray-scale imaging and then select one image for detailed quantitative interpretation by displaying value profiles. Subsequently he can interpret spatial relations of data values with 3D imaging methods. The relationships between variables and the global

structures within the data are obtained with different display modes such as multiple windows and animation. With the set of complementary imaging functions and display modes, more information is obtained than with prior imaging methods. Also, more complex simulation studies are feasible since the results can now be visualized.

In the following section, the basic concepts of data interpretation are described. Subsequently, the imaging functions are illustrated using data from large-scale simulations. Next, several examples are presented to demonstrate how the imaging functions and display modes are used for visual interpretation. The paper concludes with discussions of animation and processing steps.

The complexity of data and diversity of the applications require a display system with a range of interactive imaging and processing capabilities. In this study the display functions are implemented on a high-resolution color imaging workstation (an IBM 7350) attached to a host computer (an IBM 4341), as described in the Appendix. The workstation provides high-speed, flexible functions for data imaging and interpretation. The host computer provides data storage and management and capacity for complex computations.

Data interpretation

Before the imaging functions and their use for data interpretation are described, it is helpful to have a clear characterization of the input data and the aspects of data interpretation. The input data consist of arrays of values. Locations in the arrays correspond to positions in space, which may be a parametric space (voltage, temperature, density, etc.) as opposed to physical space. The position coordinates for each location are evaluated with appropriate scale factors. Hence, the input data are characterized as pairs of numbers, i.e., (*coordinates; value*).

Visual interpretation of the input data (coordinates; value) has three different aspects. First, for selected coordinates in the data array (points or lines in space), the corresponding values are displayed. Second, for a selected range of data values, the corresponding coordinates are displayed. Third, the distribution of values is displayed, i.e., the correspondence between coordinates and values.

Data interpretation is based on using several imaging functions to display the data in different ways and to provide complementary information. Five basic imaging functions developed in this study are the one-dimensional (1D) profile, 2D and 3D colored range, continuous tone, relief, and surface pattern. These functions are related to the three aspects of data interpretation in Table 1.

Table 1 Imaging functions for data interpretation

Aspects of Interpretation	Imaging Functions
Coordinates to values	1D profile
Values to coordinates	2D and 3D colored range
Distribution (coordinates; value)	Continuous tone Relief Surface pattern

In addition to these five imaging functions, there are two display modes on the color monitor. A very effective mode is based on multiple windows on the monitor screen presenting several images at one time.[3] One window may contain input data; a second may be a 2D color-coded image; and several windows may contain different orientations of the 3D data array. Animation is a second powerful display mode. For example, with a 3D data array, images can be displayed sequentially to present the 3D data set in different orientations. The user observes rotating structures on the screen.[3,5] If the input data correspond to the evolution of a time-dependent process, animation allows the user to view the data as a changing process.[6,7] Multiple windows can also be used with animation; different animation series are displayed in different windows.

The aggregate of the displayed images is a *hyper-image*, as illustrated in Figure 1. The hyper-image is the collection of everything the user sees. It can be thought of as an image with a time axis. Each hyper-image is described by six numbers representing color hue, lightness, and saturation, x-y screen coordinates, and time. It is a set of points in a six-dimensional space but appears to the user as a multiple-window animated color image. The advantage of this generalized approach to imaging is that color, multiple windows, and animation are used as alternate dimensions in which to display and interpret data. The basic task is to determine the best use of various dimensions in the hyper-image (windows, colors, animation) while considering the characteristics of the data and the limitations of human perception.

Visual data interpretation is based on using different imaging functions and display modes to understand the features of the data. The user develops an interpretation by interactively selecting imaging functions and modes (windows and animation series) to display the data. Several hyper-images may be required in the process of data interpretation; it is a dialogue process.

In order to describe the imaging functions and how they are used in the process of visual interpretation, data are taken from the following example of large-scale simulation. The output is three series of 3D arrays.

Description of example. As physical dimensions of solid state devices shrink, logic errors attributable to ionizing alpha particles become more frequent. The resulting flood of unwanted additional carriers within a device can cause a memory state to be changed. An alpha particle hit on an npn bipolar transistor is modeled in three dimensions using a computer simulation program called FIELDAY. The goal of the simulation is to understand the mechanism of charge redistribution and collection within the device after the hit.[3]

The 3D device geometry used for the simulation is shown in Figure 2. The emitter and base contacts are on the bottom. The subcollector contact is on the side, and the substrate contact is at the top. The alpha particle path is at an angle of nine degrees to the bottom surface as shown. The simulated region is 6.0 micrometers (μm) (x direction) by 4.82 μm (y direction) by 6.0 μm (z direction).

The internal state of the device was simulated at 16 times; the time steps are logarithmically increasing. The shortest time step was 1.0 picosecond (ps); the longest step was 5 nanoseconds (ns). The total time simulated was 10 ns. This simulation study required 1000 CPU minutes and 45 megabytes of memory, running on an IBM 3081 processor under the Multiple Virtual Storage/Extended Architecture (MVS/XA) operating system. The output of the simulation program is a data set containing the three 3D scalar arrays (electron and hole densities, and potential) at 16 time steps. The output array of electron density at 5 ps is shown in Figure 3. These 16 images correspond to 16 slices through the device parallel to the front plane in Figure 2. To interpret the output data with conventional 2D contour plots or continuous-tone images, 768 images (3 variables × 16 z planes × 16 time steps) must be mentally combined by the designer to draw a conclusion regarding device behavior. Consequently, only a limited portion of the results can be imaged and interpreted on the basis of conventional 2D imaging methods. However, with the 3D imaging and animation presented in this study, a clear display of the total carrier dynamics is obtained.

Imaging functions

Visual interpretation is based on selecting appropriate imaging function and display modes to clearly display the key features of the data. Several imaging functions are required since there are a wide range of data characteristics and different aspects to interpretation, as listed in Table 1. The five basic functions developed in this study are described in the following paragraphs. They have been used effectively in many engineering, scientific, and medical applications.[3,8,9]

1D profile. The 1D profile function allows the user to *display values* corresponding to a selected line across the data, the first aspect in Table 1. The user selects two end points with a cursor on a 2D image. The profile of values along the line is plotted, along with the selected line on the 2D image. Several profile lines can be selected and displayed in different colors. Two profiles of an electron density array are shown in Figure 4.

The 2D array corresponds to a cross section located 2.8 μm from the rear plane of the transistor in Figure 2. These profiles allow the user to compare the electron densities across the alpha particle track and adjacent to it. The density is 10^{18} per cubic centimeter in the center of the track and 10^{12} per cubic centimeter adjacent to it.

Selecting a line is a direct way of specifying locations and obtaining quantitative information. An indirect approach is to use a range of values to select locations. By using a range of values corresponding to the track, we can select points in the track. An estimate of the track volume is obtained by counting these points on the basis of the histogram of data values. A histogram of values also provides quantitative information on the global characteristics of the array.

2D and 3D colored-range imaging. The 2D and 3D colored-range imaging functions allow the user to *display coordinates* for selected values, the second aspect in Table 1. Ranges of values are selected using

1D profiles; the corresponding points are then displayed in different colors. An example for electron density is shown in Figure 5. This image corresponds to a plane that is 2.8 μm from the rear plane in Figure 2 of the device. Points whose values lie between 1.0 and 4.0×10^4 are displayed as green, between 5.0×10^{16} and 2.0×10^{18} as orange, and between 5.0×10^{19} and 10^{20} as red. The 1D profile provides quantitative information about a local portion of the data set. In contrast, the colored-range images provide qualitative information which reflects the entire data set.

A 3D colored-range image is formed by sequentially projecting 2D colored-range images.

A 3D colored-range image is formed by sequentially projecting 2D colored-range images.[8] The colored-range values of Figure 5 are used to form the 3D colored-range image of electron density in Figure 6. The rear plane is projected first. Lighter colors are used toward the front to form depth shading. The apparent orientation of the data is determined by the orientation of each plane relative to the screen and the sequential offset of the planes across the screen.

The colored-range method of 3D imaging has four advantages. First, several value ranges can be displayed with different colors. Second, as successive frames are overlaid on the CRT display, the relative position and size of various value ranges can be seen; information is obtained during the development of the final 3D image. Third, 3D images are rapidly formed since the shading is not based on computing surface normals with light-source and viewer directions. Also, the speed is independent of the options used (except for stereo, since two images are formed). A fourth key advantage of this approach is that data need not be continuous. Very irregular value regions can be displayed, as well as a disjoint collection of points.

Options for 3D imaging. A basic problem in interpreting 3D data is seeing behind near structures and inside enclosed regions. Three 3D imaging options can be used for this problem: transparency, cutout, and split. Transparency is based on displaying selected near structures with colors that are a mixture of the colors of the near and far structures. This effect is obtained by computing the required transparent color or by writing every other point of the near structure, leaving the far points to show through.[8] Transparency has limited application in data interpretation because transparency renders structures nebulous; their shape and location are more difficult to interpret.

The cutout option is a useful tool; a portion of the data set is removed to see behind or inside of structures.[3] A cutout is defined by tracing its periphery with a cursor on top of the 3D image. A portion of electron density in Figure 6 is cut out to expose the alpha track and the interior of the subcollector region; see Figure 7. The cutout depth is 4.0 μm.

The split option folds open the 3D structures along a selected line. Figure 8 is composed of the same data as Figure 6, with a split 2.0 μm from the left side and with a 30-degree opening. The advantage of a split compared to a cutout is that no parts of the structures are removed, yet the interior can be seen. A disadvantage is that its geometric shape is more limited.

Continuous-tone imaging. The third aspect of data interpretation in Table 1 is the *distribution of values.* A variety of methods have been used: small numbers or letters for values, symbols with different sizes, varying dot densities, lightness of each point, and a 3D mesh surface.[1,3,6,10] For engineering, scientific, and medical applications, three methods are useful for displaying the distribution of values: continuous-tone imaging, relief imaging, and surface pattern imaging. The simplest mode of displaying the data is a 2D continuous-tone image like that used in the 1D profile (Figure 4). The array values can be displayed in gray levels or in a range of colors. By using a few gray levels or colors (10 to 15), contour lines are obtained. Contours yield quantitative information and are especially informative when used with animation to display changes.

Relief imaging. Relief images are based on displaying the 2D array of a number as a surface in three dimensions whose height is proportional to the array values. Surface points with different heights are dis-

played with different colors; higher points are displayed with lighter or brighter colors. A relief image provides a better interpretation of amplitude than that obtained with a continuous-tone image. It requires a few seconds to be formed, whereas continuous-tone images are displayed rapidly. Figure 9 is a relief image of the same data used in Figure 4. Different projection directions can be used to see behind large peaks.

Surface pattern imaging. A surface pattern image is similar to the relief image in that the data are displayed as a surface in three dimensions. The difference is that a pattern is placed on the surface, and it is viewed from the top. Depth is displayed by changes in the surface pattern shapes, sizes, and colors. Continuous-tone and relief imaging can be used for any type of data, but are less effective than a surface pattern for slowly varying data. Surface pattern imaging is illustrated in Figure 10. The input data consist of a 2D cosine function with changing period. With the surface pattern it is easy to compare the slopes of the three bands and the gradual change in the upper left corner.

Examples of visual interpretation

To demonstrate how the imaging functions are used with multiple windows and animation, several examples are presented. The basic task is to select an imaging function and the best use of various dimensions in the hyper-image (windows, colors, animation); see Figure 1. Selecting the correct hyper-image can very significantly improve interpretation. This is demonstrated by comparing several hyper-images of the same data, i.e., simulation data of an alpha track in an npn transistor. The problem is to relate the track to other structures in the device.

First, different hyper-images are compared using the continuous-tone imaging function. A multiple-window display with black-and-white images is not very effective; see Figure 3. It is difficult to interrelate 16 images; the continuous gray scale does not highlight the location of the track. With animation instead of multiple windows, the relative position of the alpha track is more clearly seen, since it is the only structure that changes between frames. The time dimension of the hyper-image is used instead of the space dimension. It is not possible to present the full power of animation with a static image on a printed page. But some appreciation can be obtained from six frames of the animation series in Figure 11A.

Since visual interpretation with animation is based on observing edge motion, sharp edges are more easily followed than nebulous edges. By using 20 gray levels to obtain contour lines, the position of the track is made more apparent in animation, as in Figure 11B. This is done by simply using different coordinates on the color axes of the hyper-image. Finally, with color, a very clear indication of the track is obtained, as in Figure 11C. Figure 3 and Figure 11C are based on the same data and use the same imaging function, but the appropriate hyper-image significantly clarifies the location of the track.

The images in Figure 11 are different hyper-images based on the same imaging function. Using the 3D colored-range function, the input data (Figure 3) can be displayed as a single image; see Figure 6. However, front portions of structures obstruct the view of rear portions. With use of the cutout option, an animated series of 3D images with different data orientations clarifies the location of the track in relation to other structures, as seen in Figure 12.

Animation provides very powerful position and shape clues.

Animation provides very powerful position and shape clues that cannot be obtained from a set of static images displayed in multiple windows. Interpretation is significantly enhanced by using the time axis of the hyper-image instead of the spatial axes.

The 3D animation series in Figure 12 does not replace the 2D color animation series in Figure 11C. They are different hyper-images, and both provide useful information for data interpretation. Figure 11 displays the *distribution of values* (coordinates; values) in Table 1; no data are omitted. Figure 12 displays the *locations* of key ranges of electron densities. Various imaging functions and display modes provide complementary information for interpretation.

To illustrate additional hyper-images, a second example is used. The input data comprise 100 2D arrays which represent the evolution of a chemical

Figure 1 Data display structure

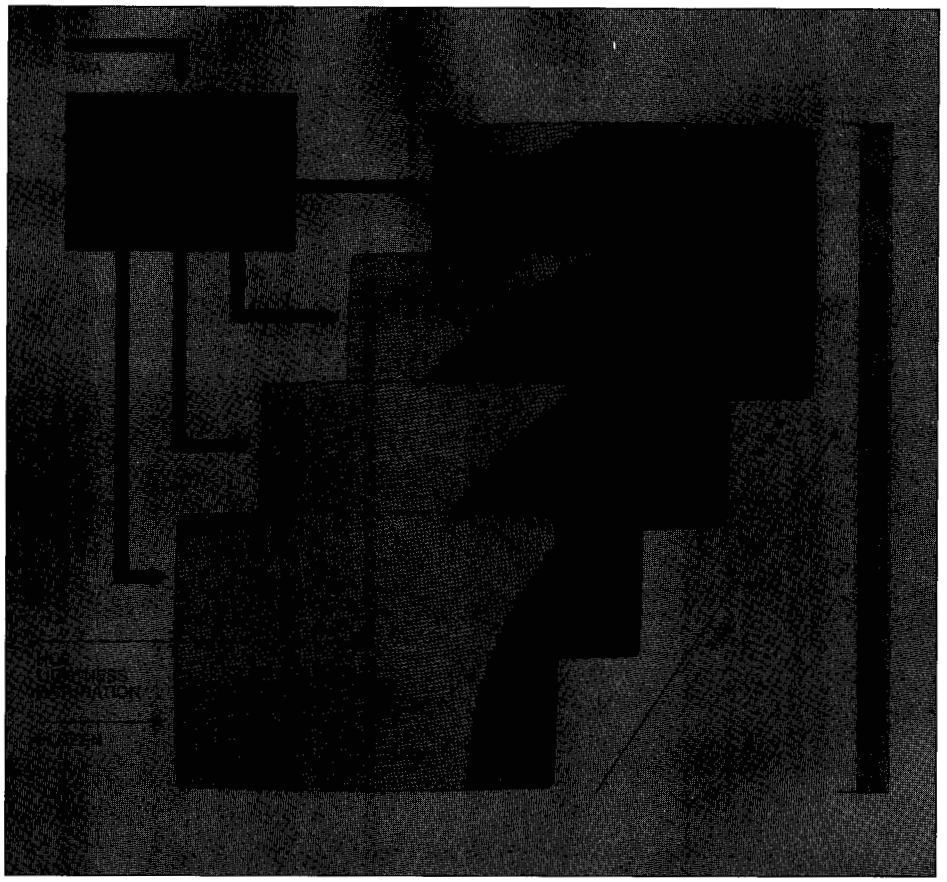
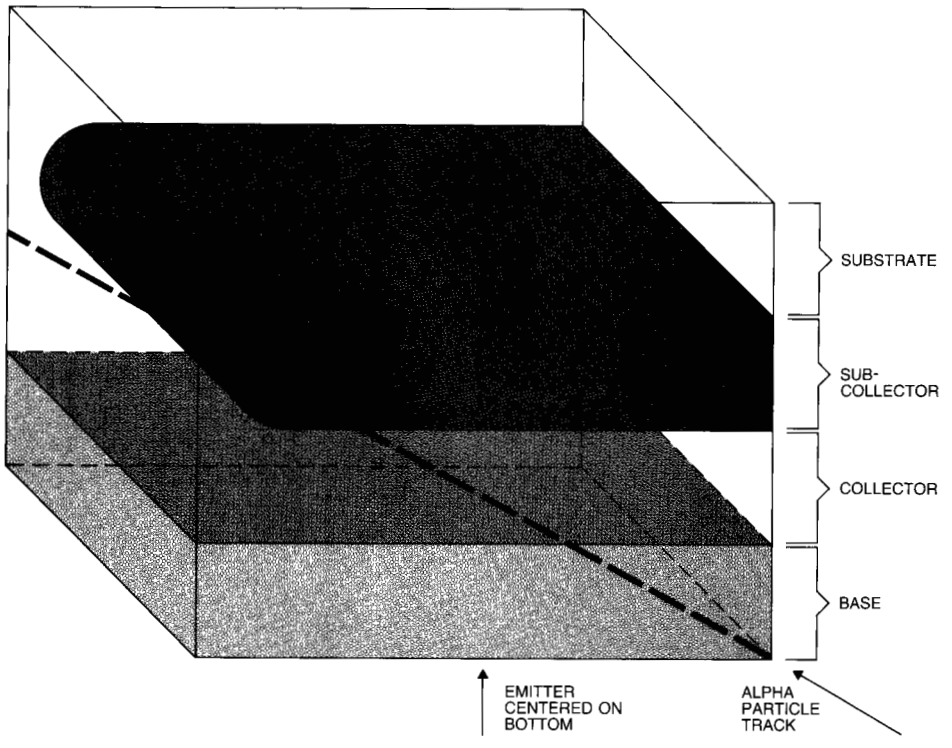


Figure 2 Diagram of simulated bipolar npn transistor



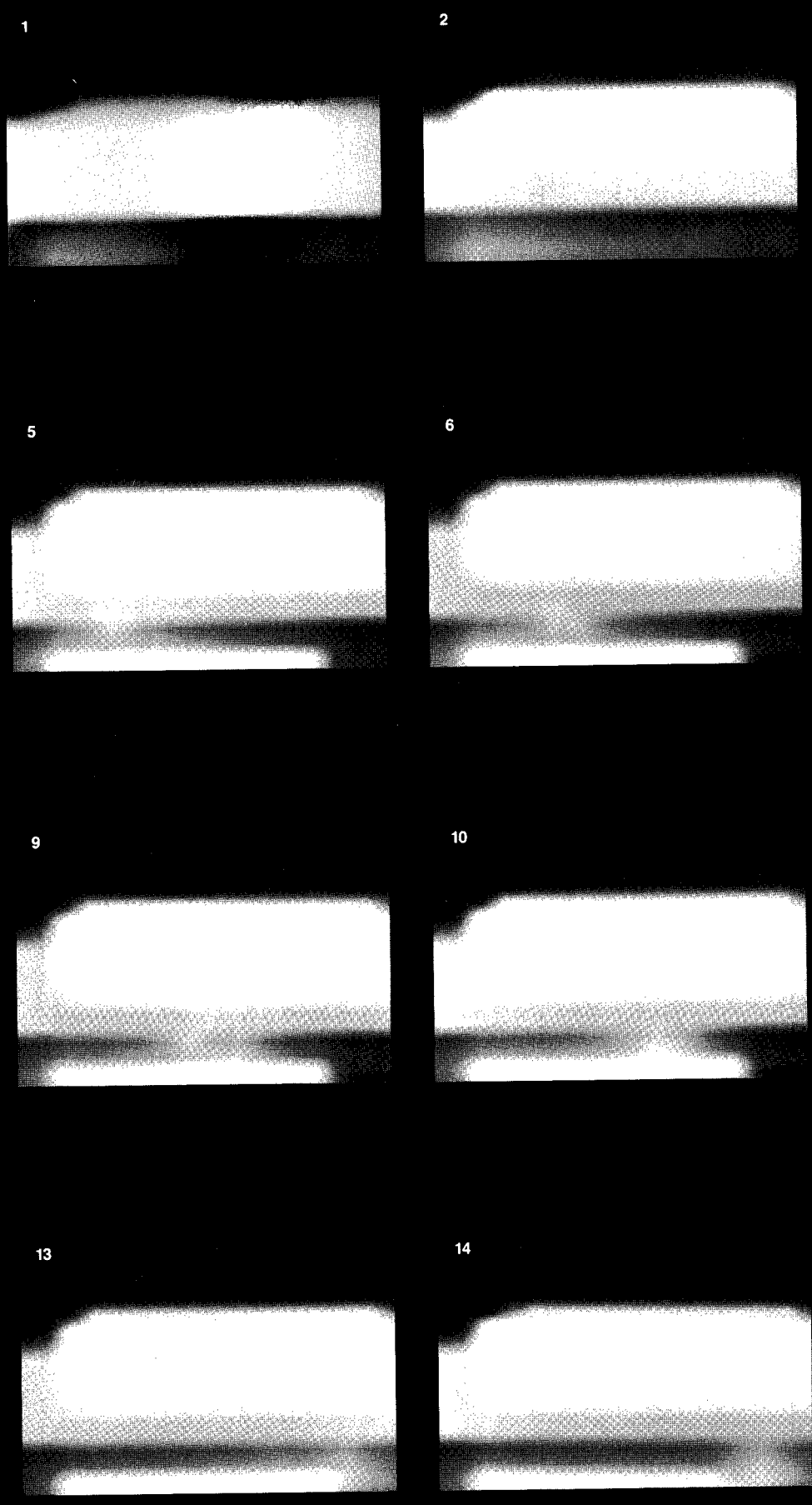
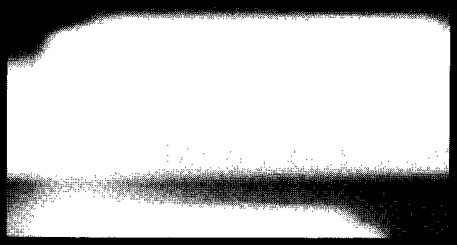
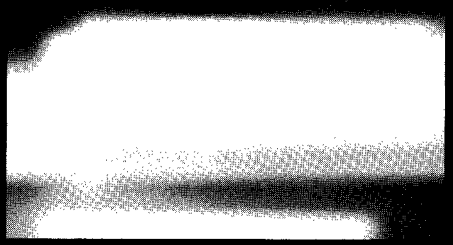


Figure 3 Electron density in an npn transistor 5 ps after an alpha particle hit

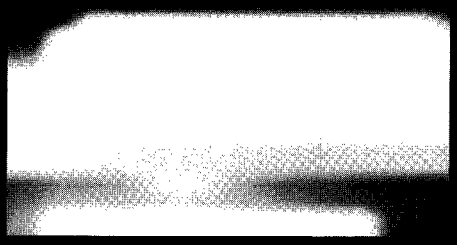
3



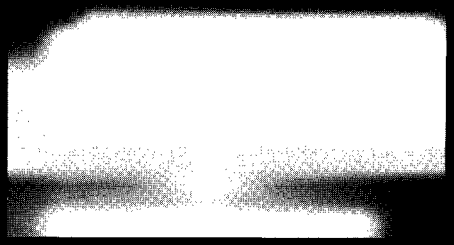
4



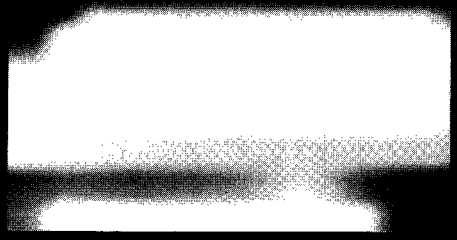
7



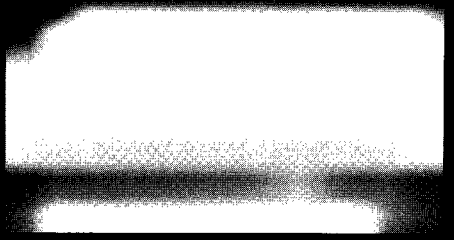
8



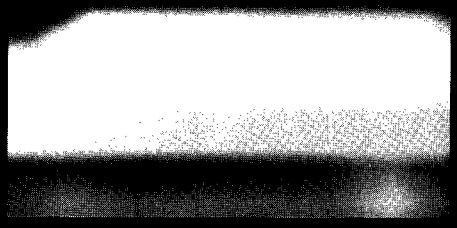
11



12



15



16

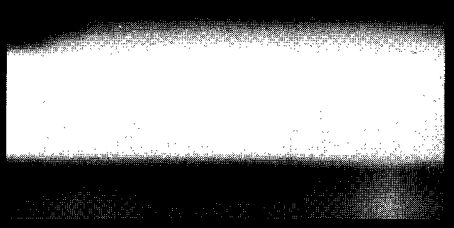
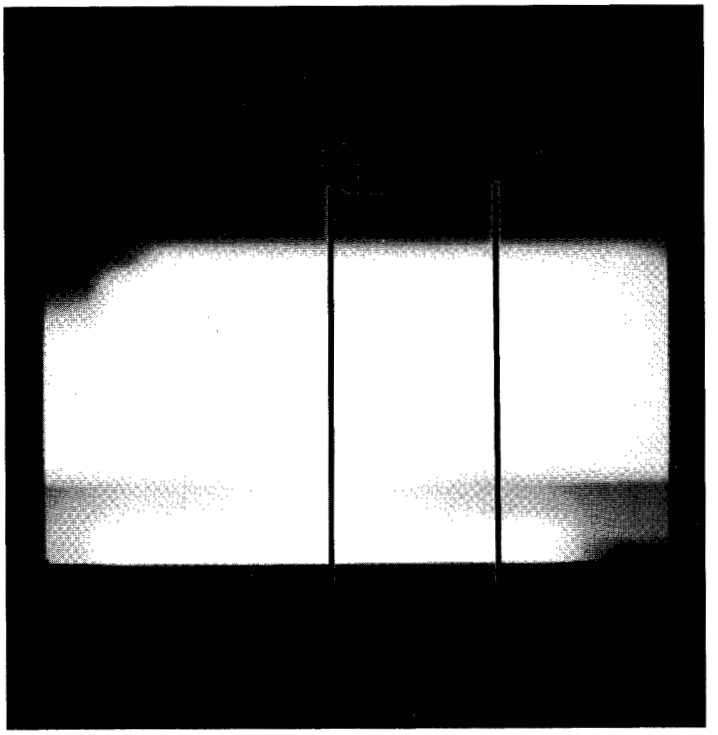


Figure 4 1D profiles from npn transistor cross section



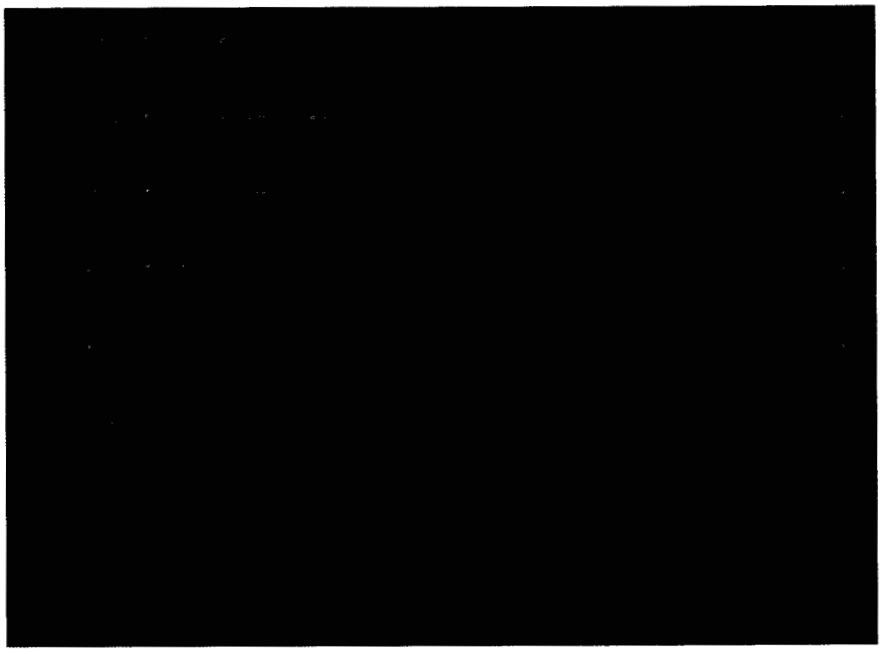


Figure 5 2D colored-range image of transistor cross section

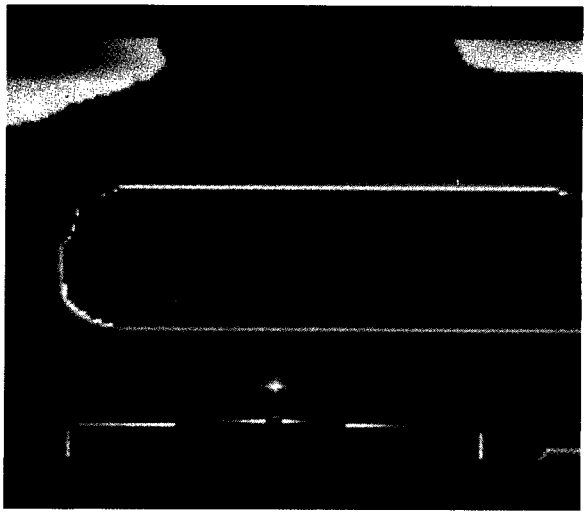


Figure 6 3D colored-range image of transistor with alpha particle track

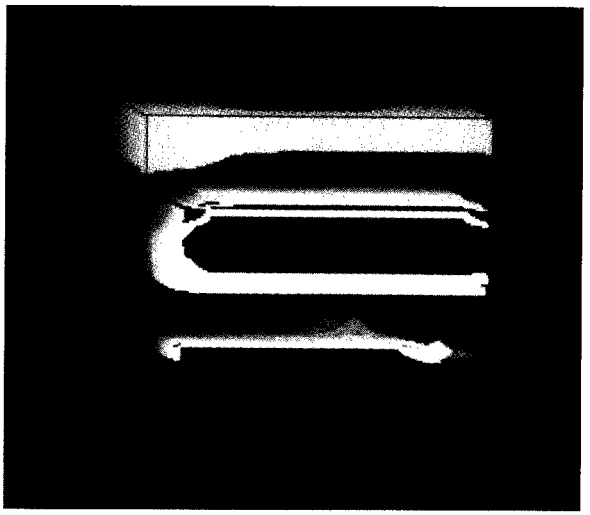


Figure 7 3D colored-range image of transistor using cutout option

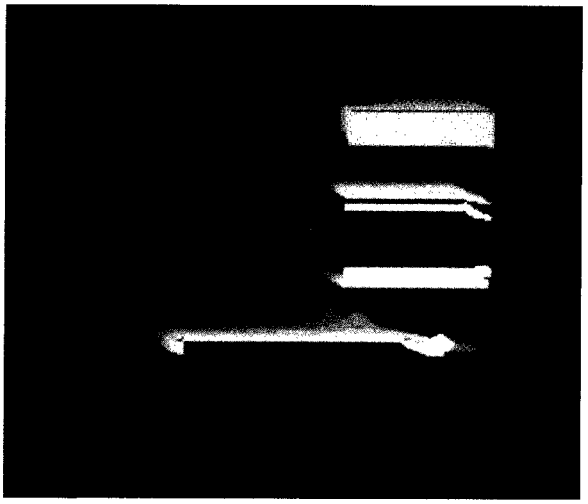


Figure 8 3D colored-range image of transistor using split option

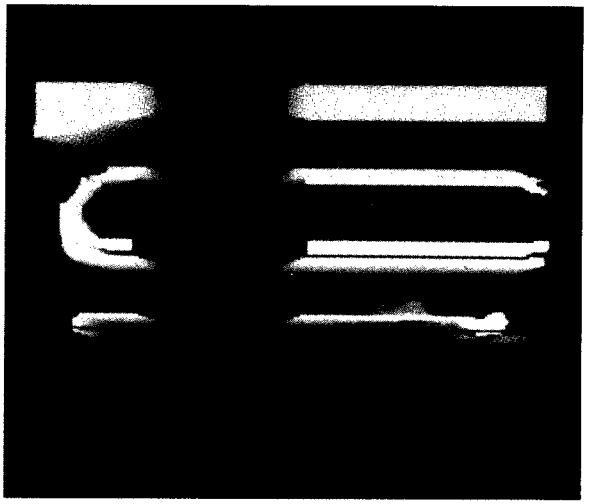


Figure 9 2D relief image of transistor cross section

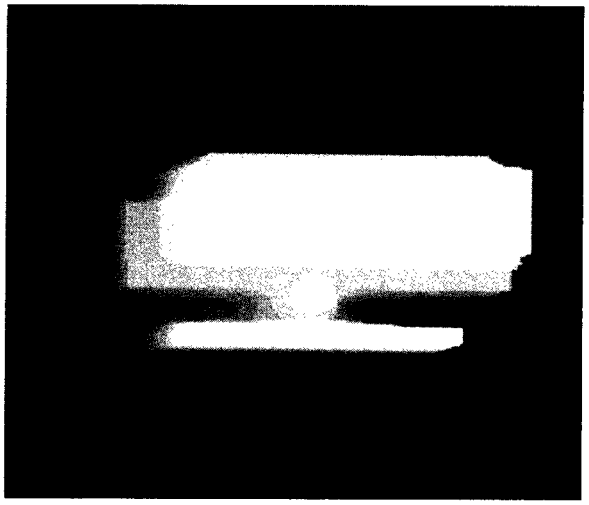


Figure 10 Surface pattern for a damped 2D cosine function

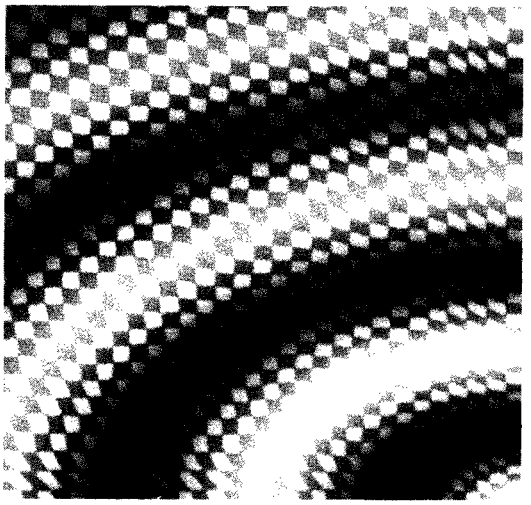
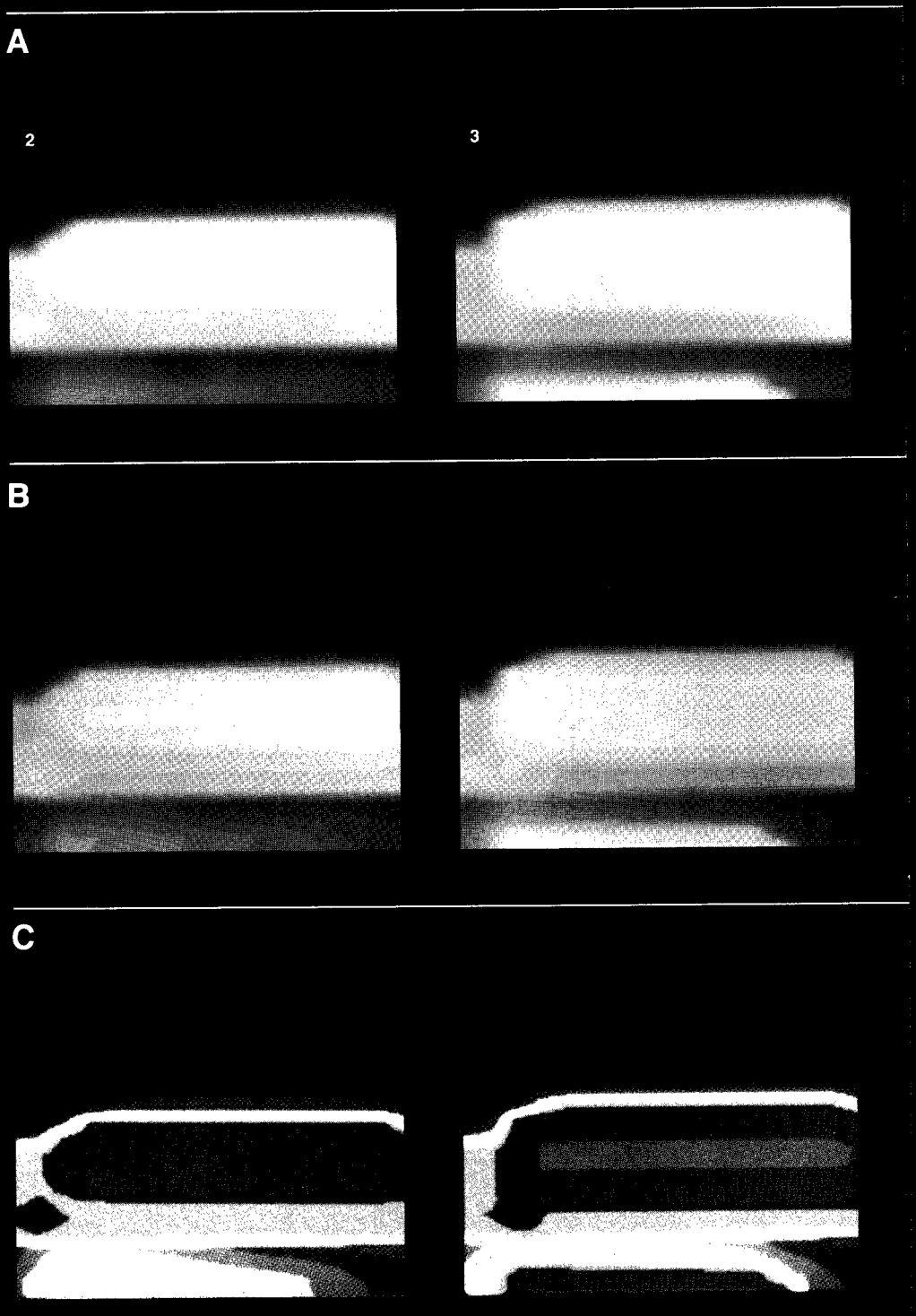


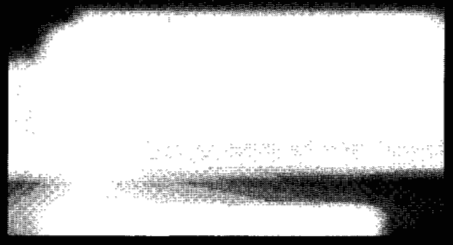
Figure 11 Animation series of electron density cross sections: (A) Continuous-tone black and white images over space; (B) Black and white images with contours; (C) Color images with contours



Continued on next page

4

5



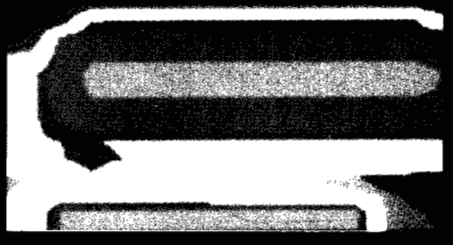
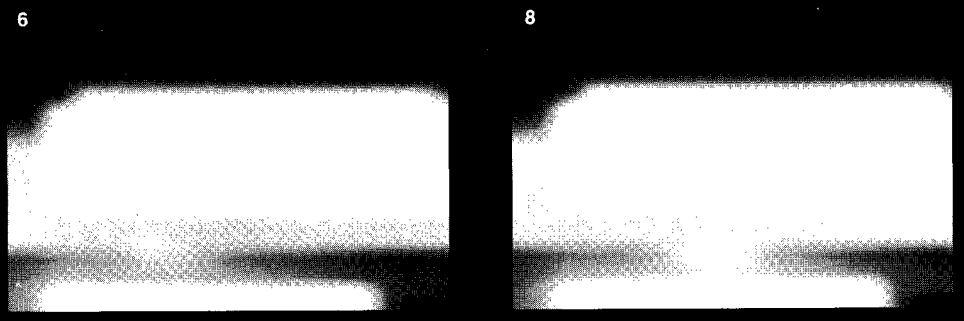


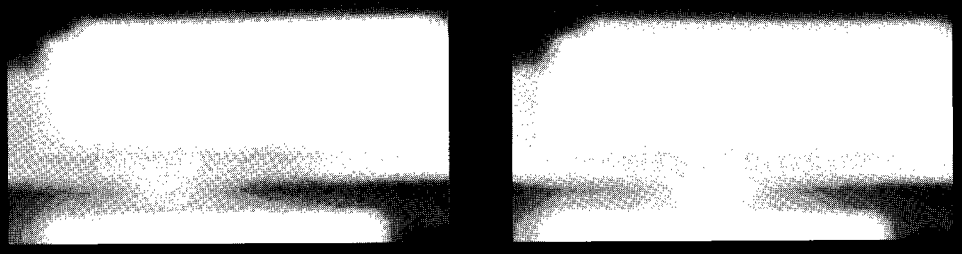
Figure 11 Animation series of electron density cross sections: (A) Continuous-tone black and white images over space; (B) Black and white images with contours; (C) Color images with contours

Continued from previous page

A



B



C

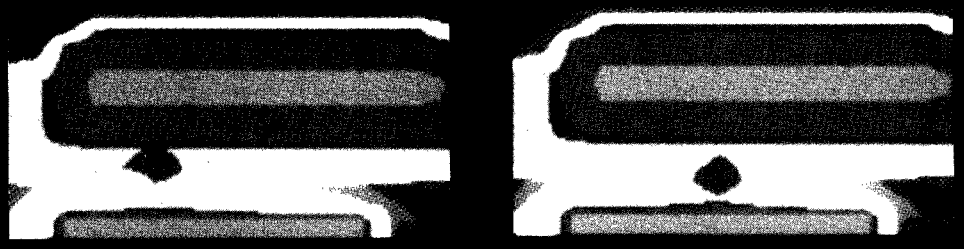
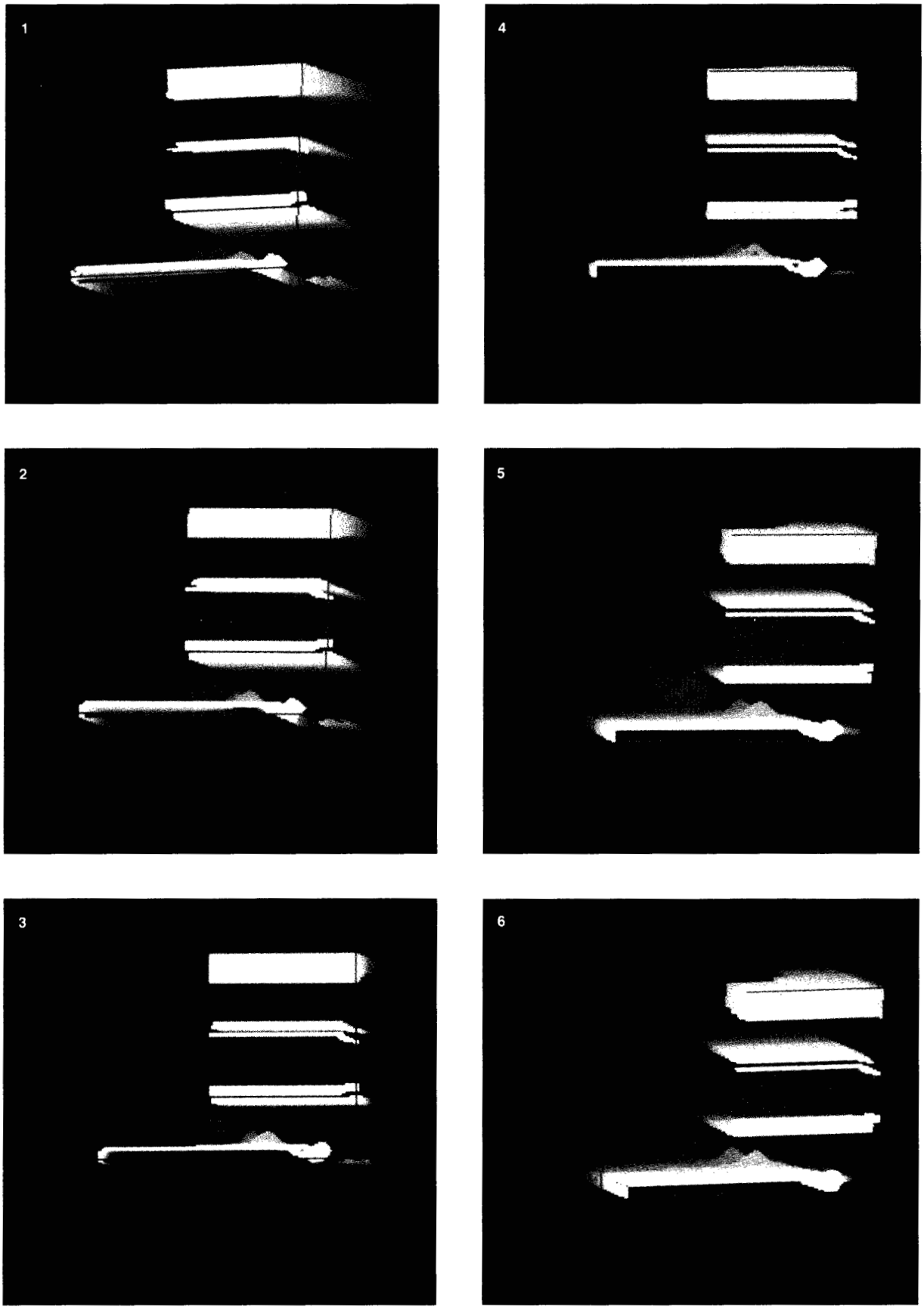


Figure 12 3D images of electron density at different orientations



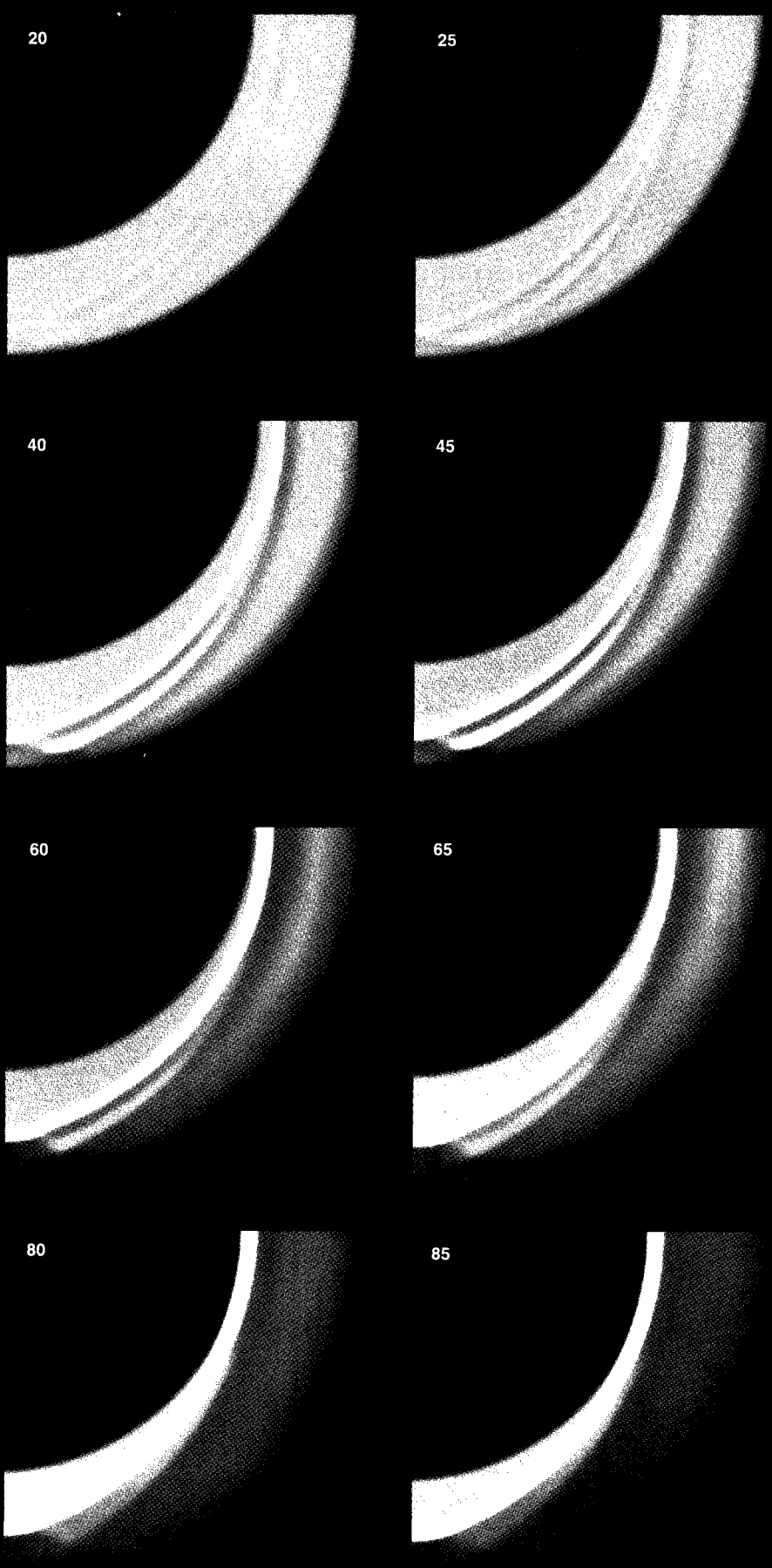


Figure 13 Mass density data over time

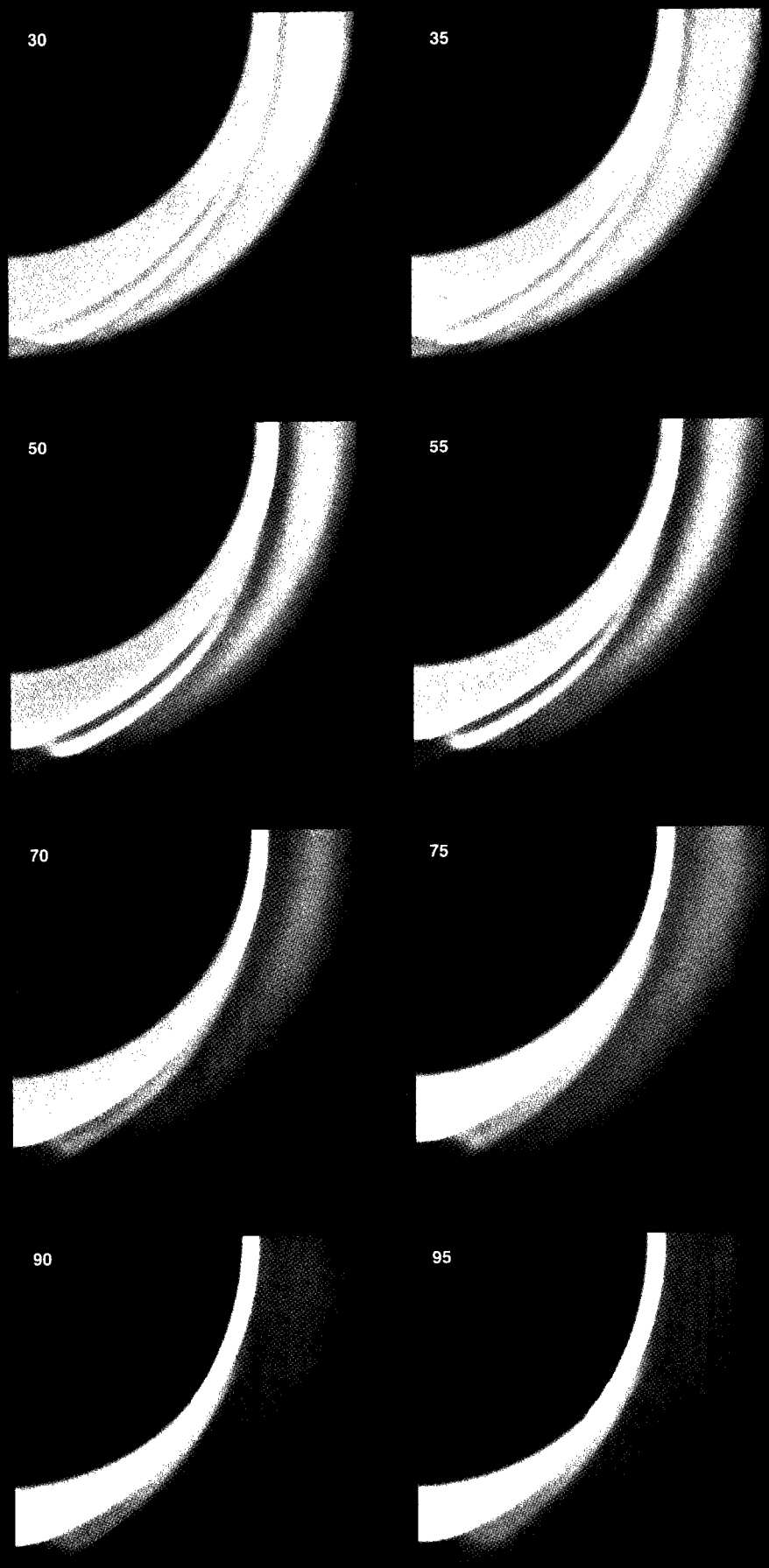
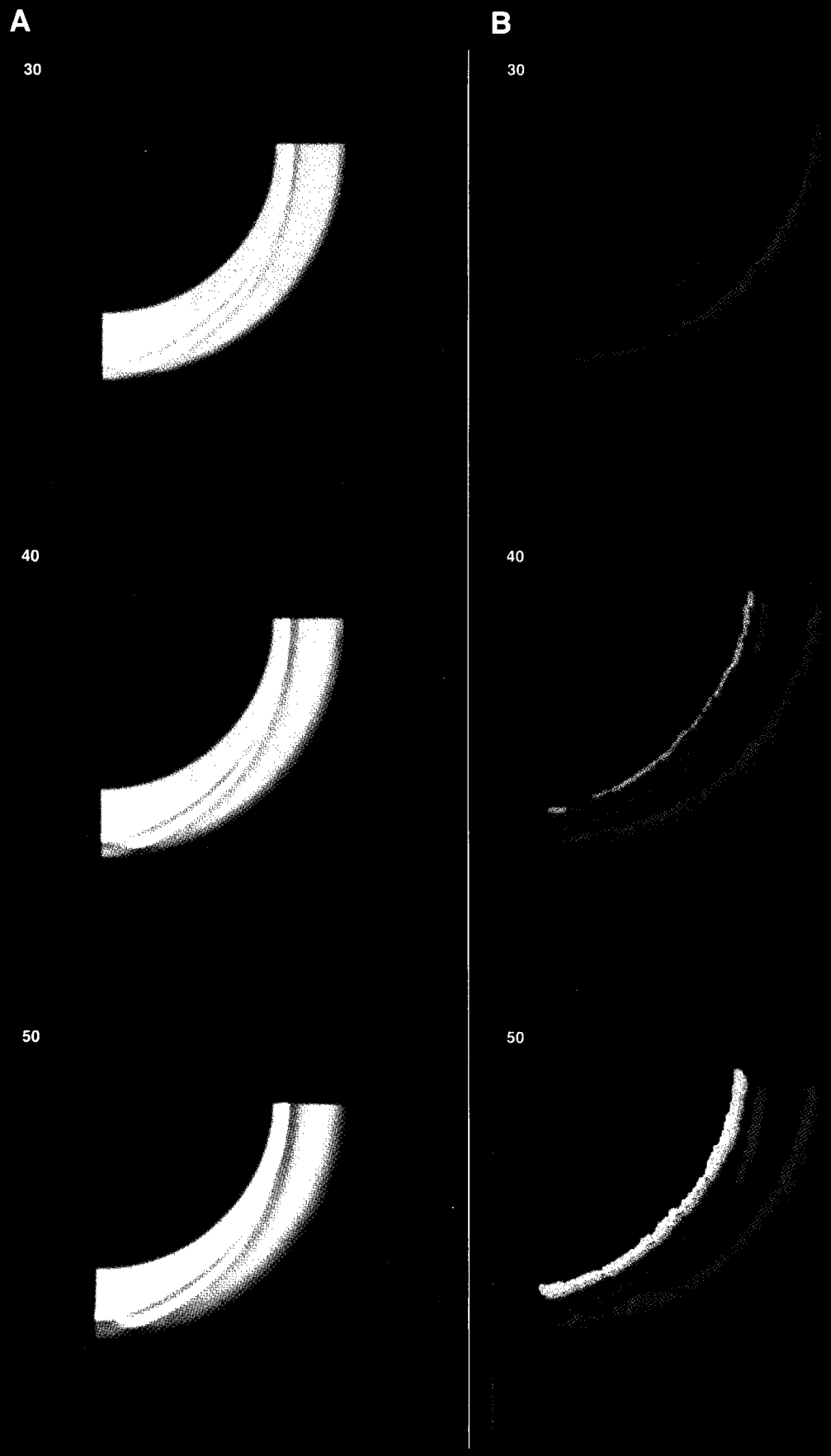

30
35
50
55
70
75
90
95

Figure 14 Mass density evolution, comparison of hyper-images: (A) Continuous-tone images; (B) Color relief images



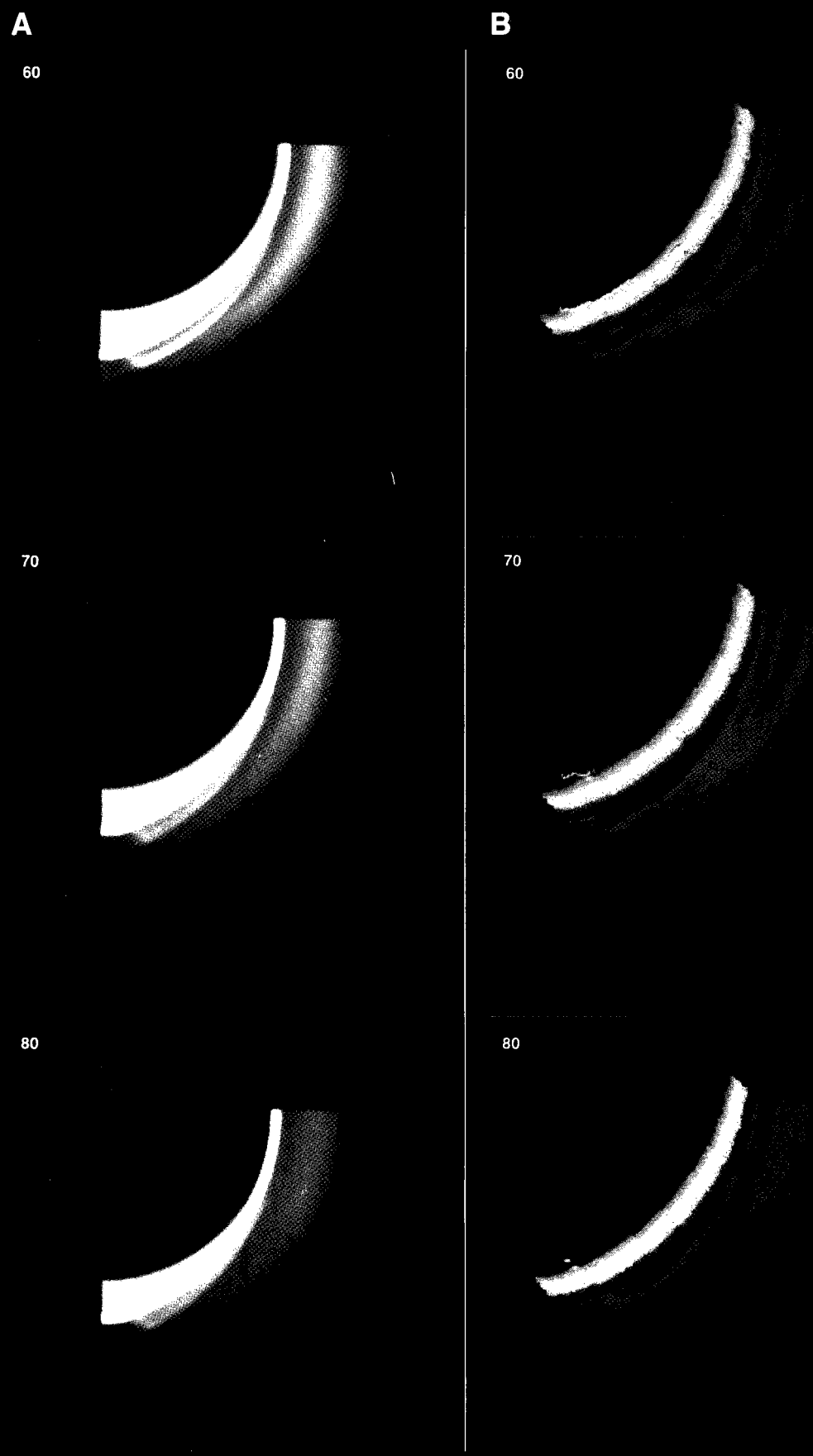
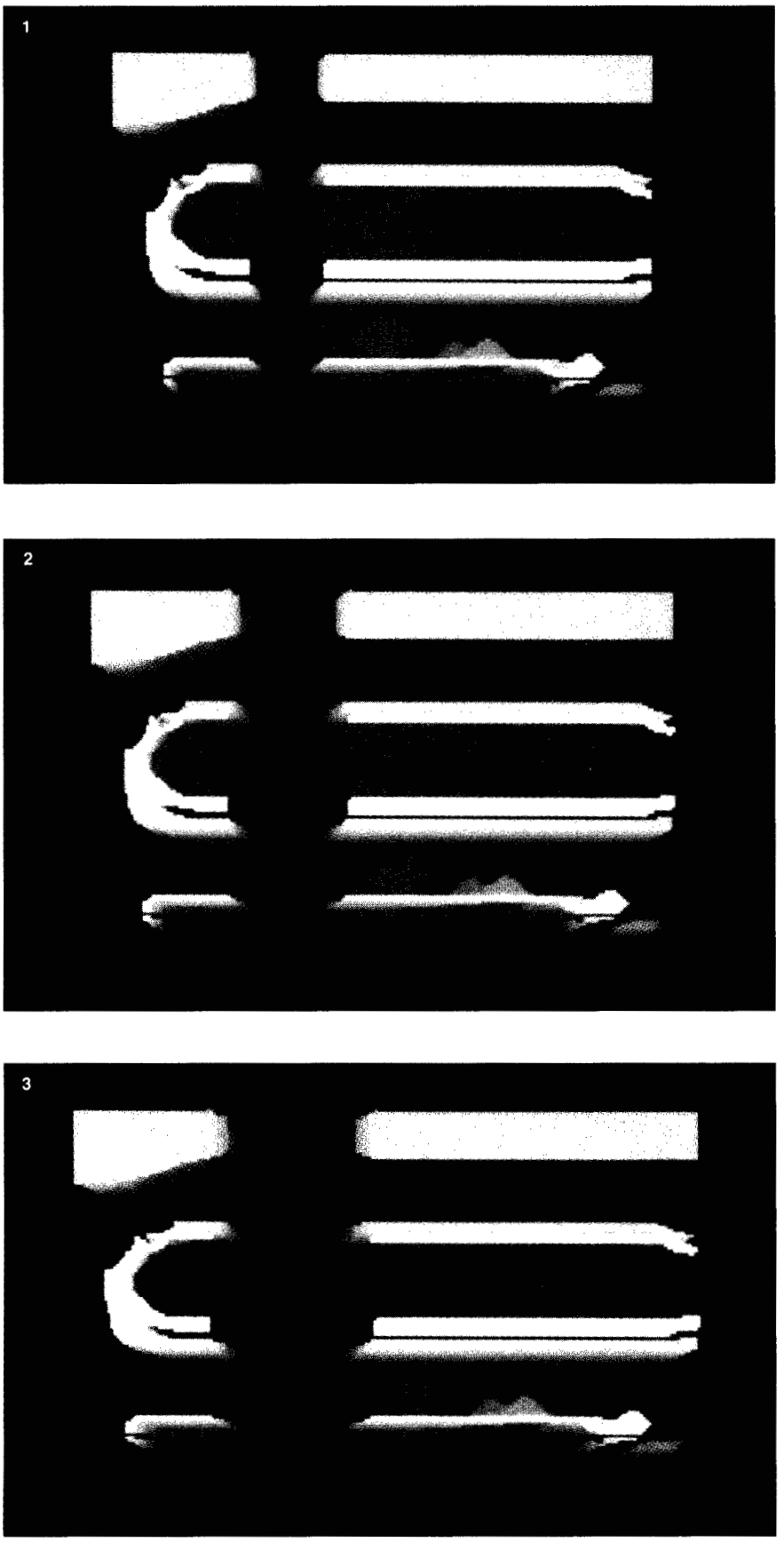
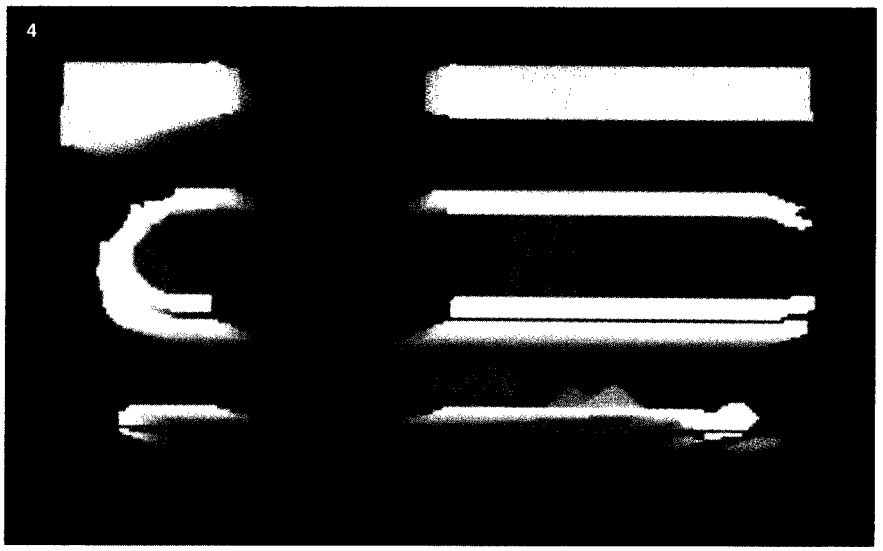
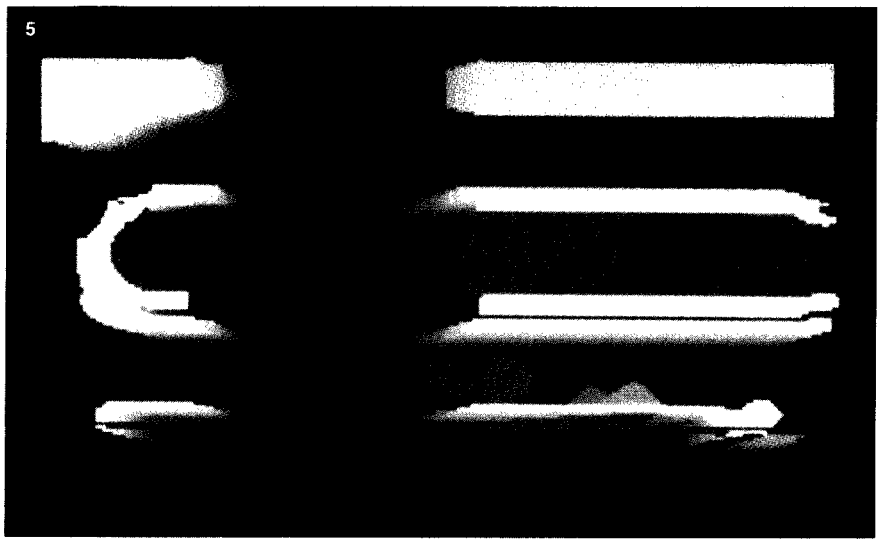

A
B
60
70
80
60
70
80

Figure 15 Animation series of electron density with increasing split angle




4


5

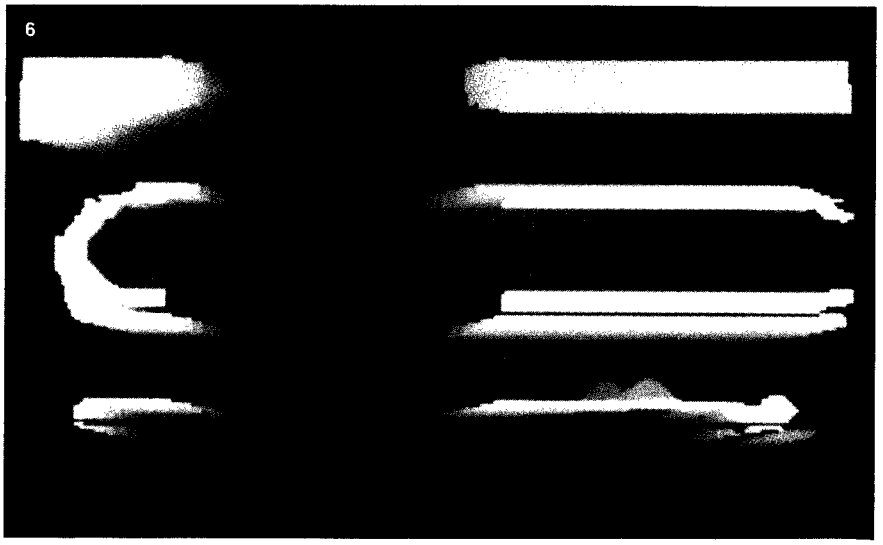
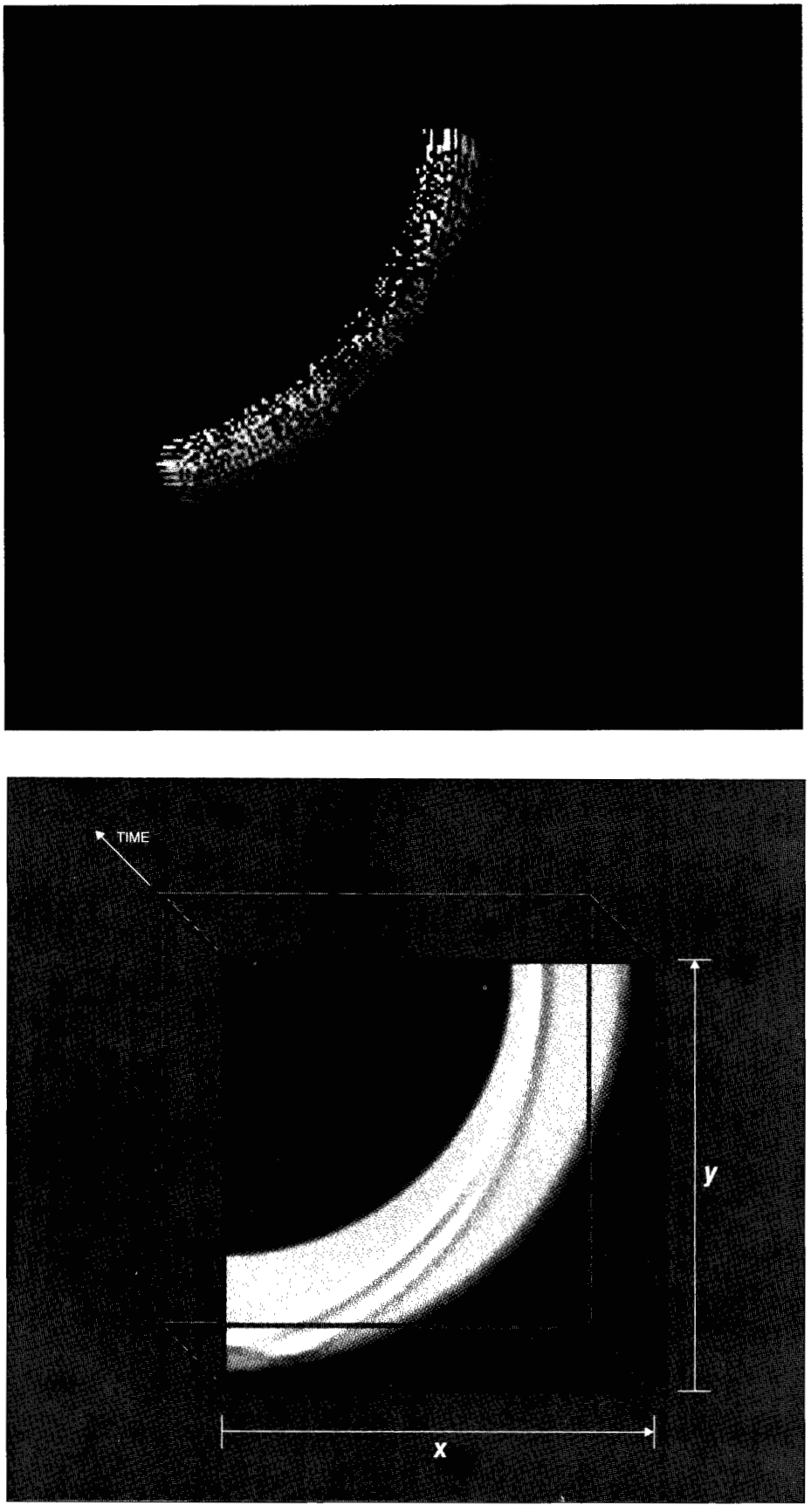

6

Figure 16 Display of high-mass density location using a 3D image and graphical description of axes



process; each array is 128 by 128 computed values of mass density. Figure 13 is part of the input data, every fifth frame starting with 20.

The distribution of values can be displayed similarly to the electron density data in Figure 11. However, for this application an animation series of color relief images provides a clearer display. A series of six input images and corresponding 2D relief images is shown in Figure 14.

The color range for the relief amplitude is chosen so that large density values are displayed in red; they are easily distinguished from the lower values, which are displayed in green and blue. As the animation series evolves on the screen, regions rise as the density increases, and the peaks change color as they rise. Also, the width at the base increases. The appropriate hyper-image significantly enhances interpretation (compare Figure 14 to Figure 13).

The 3D imaging function is useful in interpreting the *location* of high-density regions, as shown in Figure 16. In this example, the location is the position in the x-y plane and the time is along the z axis. This single image provides the user with several important results. It shows when and where the high density occurs, and it also displays the nonuniformity of the process over time. The wave-like patterns visible in Figure 5 are not seen in the 2D images in Figure 13. These examples demonstrate that data interpretation is an interactive process; each hyper-image provides additional information for interpretation.

Multiple display windows can be used in several ways. In the above examples, the same data and imaging functions were used in each window. However, different data can also be displayed in different windows.[3] For example, with the alpha particle simulation, different windows can be used to display electron density, hole density, electrical potential, and doping with the 3D imaging function. With animation, each window presents the redistribution of these quantities with time following the hit. (The doping is a fixed reference image.) Also, different imaging functions can be used in different windows. Input data can be displayed in one window with the continuous-tone imaging function, a rotating 3D image of the data can be displayed in a second window using animation, etc.

From the interpretation of a wide range of engineering, scientific, and medical data sets, certain guidelines have been observed for hyper-images. For selecting windows, using more than four display partitions does not significantly reduce the time needed for data interpretation. For example, it is more effective to have a series of four screens each with four image windows than to have one screen image with 16 image windows. For animation, 10 to 20 images are often adequate when distributed over a range of images. For selecting color ranges, more than three colored ranges confuses the 3D image; two or three is preferable.

Animation

Animation can be used in several ways to aid data interpretation. An animation series of 2D relief images can be used to display the evolution of a process like that in Figure 14. Changes in the location and magnitude of mass density are easily seen. In other applications the user may wish to interpret changes in the position and shape of certain regions in a series of 2D arrays based on continuous-tone images. With animation, coherent motion of regions is easily seen against a noisy background. The effect of frame-to-frame incoherent noise is significantly reduced. In some cases it is possible to see structures that are not apparent in individual frames viewed statically. The series may correspond to changes in time, physical locations, or a parametric variable.

Animation can be used to form a dynamic 3D display for improved interpretation of spatial location and shape. For example, if 3D images with different orientations (Figure 12) are rapidly displayed, the user observes rotating 3D structures. Spatial perception is improved in three ways. First, as the collection of data rotates about its center in space, points in the front and back can be easily distinguished since they move in opposite directions in the 2D image seen by the viewer. This provides very powerful depth clues. Second, a clearer perception of 3D shapes is obtained since structures are seen from several different directions. Third, disjoint fragments of the same structure are seen as parts of the same structure since they move together in a coherent manner.

A second example of dynamic 3D display is based on the split option; an animation series is formed with an increasing split angle. For example, Figure 15 is part of an animation series of electron density distribution. The split is located 2.0 μm from the left side, and the split angle ranges from 10.0 to 40.0 degrees. This animation series allows the user to see

inside convoluted structures as they open and close with animation and to interpret interrelations in a manner no other function allows.

Animation is also a valuable tool for system design when used in conjunction with large-scale computations. For example, Curtis and Schwieder use computer-animated graphics in the analysis of nuclear power reactors.[11] The display is an animated diagram of a simulated system, with various variables and functions represented by graphical icons. Garbini, Lembersky, Chi, and Hehnen apply similar graphical animation methods in the design of a log-processing facility.[12] Calkins and Ishimaru use animation to interpret the dynamic response of ship designs in a simulated sea;[13] vector imaging methods are used. Farrell, Laux, Corson, and Buturla describe how 3D animation can be used to interpret the response of a transistor design to an alpha particle hit on the basis of large-scale computations.[3] The evolution of the 3D distribution of hole density, electron density, and potential are displayed with 3D colored-range images.

Processing steps

Visual interpretation involves several data processing and imaging operations. The processing steps and how they are implemented reflect the system structure, which is described in the Appendix. Processing is divided into seven steps:

1. Data retrieval
2. Reformatting
3. Preprocessing
4. Qualitative review
5. Data interpretation dialogue
6. Color and shading selection
7. Image save

To display and contrast several data sets, the user must have an effective *data retrieval* system for large data sets. Typically an input data set contains two to fifty megabytes of data. Because a limited amount of data can be retained in the workstation or active host memory, data must be stored elsewhere. Several modes of data storage are available: tape, disk, memory, and workstation buffer storage. The modes range from slow-access, large-volume storage to rapid-access, intermediate-volume storage. The distribution of data among different storage modes depends on the application requirements and data volume, and requires careful consideration.

Input data can be obtained from experimental measurements, scanning instruments, computations, or simulations, and may have different *formats*. For example, the output tapes of certain medical computerized tomography (CT) scanners use an image-map algorithm to compact the data. Large-scale digital simulations often use a computation mesh with irregularly spaced points. The data must be interpolated into a rectilinear grid before display with a raster system. As part of the reformatting, it may be necessary to rescale the data. Simulation results on solid state devices have a range of 20 orders of magnitude. Such data must be logarithmically scaled to one byte prior to imaging.

A color display has greater information capacity than a monochromatic display.

After reformatting the retrieved data sets, it may be appropriate to *preprocess* the data in the host before loading the data into the workstation. The user can smooth the data and can expand or compress the data with linear interpolation. In some cases the region of interest lies in a limited portion of the input data arrays. The time needed to display images is significantly reduced by selecting a subimage. If a series of 2D images is the result of scanner measurements, normalization may be required to make the images comparable. The normalization may involve position corrections and amplitude normalization based on a histogram of data values. For medical CT scans it is sometimes useful to resample the data. The input scans are normally taken perpendicular to the body axis. Resampling is an interpolation of the input data to form a series of images as if the scanning were done along a different axis.

A rapid *qualitative review* of the data is the first step after preprocessing and loading the data into the workstation. It is necessary to confirm that the correct data were loaded and that the preprocessing did not distort the data or introduce any artifacts. Simple black-and-white images are satisfactory.

Data interpretation is a dialogue process in which the user applies a series of imaging functions and display modes to extract key features in the data. Flexibility, speed, and usability are the primary system requirements. Usability requires entry panels which express parameters in terms of the user's application. Shadow and surface reflection realism for 3D imaging are far less important considerations.

A color display has greater information capacity than a monochromatic display. Each display point has three attributes with a color display (hue, lightness, and saturation) and only one attribute with a monochromatic display (lightness). However, color must be selected carefully for a clear and accurate display of the data. An important step in data processing is to select *colors and shading* that enhance the structures of interest. A perceptual color space can be selected in which uniform increments in hue, lightness, and saturation result in uniform increments in perceived color.[14–19] The objective is to select a fixed set of colors that is effective for a wide range of applications. However, the visibility of a structure is influenced not only by its colors but also by the surrounding colors.[20,21] A fixed set of colors is useful for a limited range of images. An effective method of selecting the color for a structure is to modify its color while viewing the entire image.[8] For example, an x-deflection of a joystick changes the hue of the selected structures, and a y-deflection changes the lightness. Color saturation is less influential in imaging and is entered via the keyboard.

The user may wish to *save* a hyper-image for redisplay, for viewing without an imaging system, or for publication. Results can be saved in softcopy on the host disk or on magnetic tape, or in hardcopy as photographic prints, slides, 16-mm film, or videotape. The Matrix Color Graphics Camera is used for the prints in this paper. Animated images are recorded directly from the screen with a video or 16-mm camera electronically synchronized with the color monitor. Videotape is a convenient method of saving an animated image; however, only low-resolution images can be saved with standard recorders. Sixteen-millimeter film is superior in resolution, with a greater color range.

Concluding discussion

Computer-aided mechanical design and data interpretation both form images of 3D structures, but they involve fundamentally different approaches. Computer-aided design is based on the display and manipulation of known geometric structures. Quantitative geometric relationships are used to form an image; for example, a cylinder is described in terms of its diameter, height, position, and orientation. Data interpretation is the converse of computer-aided design. The geometric relationships within the data are unknown; an image is used to characterize the geometric relations. This basic difference between computer-aided design and data interpretation results in different data storage methods, image formation algorithms, and structure manipulation. Data interpretation requires a flexible, interactive display system for a large volume of data.

A key aspect of data interpretation is the approach to 3D display. There are two basic approaches: electro-optical and digital. Electro-optical approaches include varifocal mirrors, rotating multidiode arrays, and holography.[22,23] In general, these systems are designed to display 3D structures and do not support the range of imaging methods and interactivity required to interpret complex multivariable data. A digital system consisting of a host computer and an image processing workstation provides flexible, interactive imaging that is well suited to data interpretation. Further, recent hardware developments provide improved CRT resolution, low-cost storage in the workstation, and high-speed special-purpose hardware for image processing.

The perception of 3D structures from a 2D screen image is based on the visual clues humans use to interpret the spatial relationships of objects around them, such as shading, colors, stereoscopy, relative motion, atmospheric and geometric perspective, and surface texture.[24,25] Spatial relationships can be perceived with incomplete or inaccurate visual clues, which is very useful in a digital display system. A 3D object can be perceived using simple models for shading, perspective, motion, etc. Often simple models have implementations that are rapid and interactive. For example, a series of closely spaced points are perceived as a line because of visual closure.[25] The reference frame in Figure 15 is a series of separate colored points. Stereoscopy is perceived with almost any two images of an object at different orientations; stereoscopic perception is insensitive to how the rotations are formed by the computer. Continuous motion is perceived with a discrete series of images displayed rapidly, based on the phi phenomenon.[25] The perception of shape based on surface texture is the basis of the surface pattern display in Figure 10. Surface contour lines also provide shape and size clues.[26] Stroebel, Todd, and Zakia describe several of these principles of perception in more detail.[25]

The methods used to display 3D scalar data are based on visual clues used in everyday experience. Points corresponding to a specified range of values are displayed as a 3D structure, as described earlier. In contrast, there are no simple visual clues for displaying vector fields. Flows and gradients are not seen directly; structures are. Also, more data must be presented to the user for his interpretation. In the case of 3D vector fields, three times the volume of data must be displayed compared to 3D scalar fields.

The methods used to display 3D scalar data are based on visual clues used in everyday experience.

For 2D vector fields, small arrows are useful for simple, slowly varying fields.[4] Arrows are impractical for 3D vector fields. An alternate approach is to display the magnitude of the field using the methods described above and to overlay direction lines. This approach is used by Farrell, Laux, Corson, and Buturla to display current flux in a transistor cross section.[3] An advantage of this approach is that the direction lines and magnitude field can be displayed with other variables. For example, the current flux lines can be displayed with the electrical potential. This approach can also be used to display 3D vector fields, provided stereoscopy or animation are used to clarify the location of the lines in space.[3] Direction lines require special computations since simulations result in vectors at a set of mesh points. Smooth lines must be formed to match the direction of the vector field at the mesh points. In electromagnetic applications there is an additional consideration—there may be sources or sinks of flux lines. Imaging 3D vector fields is an important problem for future research.

Visual interpretation of vector and scalar data is an essential part of many large-scale simulations and computations. Simple display methods based on contour plots, 3D wire-frame images, or black-and-white halftone images do not have sufficient capacity. The set of imaging functions and display modes used in this study are a powerful set of complementary tools for a range of applications in engineering, science, and medicine. With the appropriate hyperimage, interpretation is significantly improved; as an example compare Figure 3 and Figure 12. Further, more complex simulation studies can be conducted since these methods facilitate interpretation. For example, the simulation of an alpha hit generates three 3D arrays (electron density, hole density, electrical potential) over a series of time steps. The simulation results in several hundred 2D images which cannot be displayed, let alone interpreted, without 3D imaging and animation. Also, these imaging methods can be used during the course of long computations to evaluate the distribution of numerical errors and numerical convergence.[3]

Acknowledgment

During the course of this research, discussions with C. N. Liu and R. H. Wolfe of IBM and R. A. Zappulla of Mount Sinai Medical Center have been very helpful. Their generous assistance has improved the imaging functions and the usability of the system for data interpretation. The transistor simulation data were kindly provided by P. Corson of IBM, and the mass density data were provided by K. Winkelmann of IBM and G.S.I. West Germany.

Appendix: Image processing system

The image processing system has two basic components, the host computer and the image processing workstation. The host performs data file management, selection of subimages, and initial data smoothing. The workstation provides interactive 3D imaging with rotation, data cutout, smoothing, split, and color selection. For this study the host computer is an IBM 4341. The workstation is an IBM 7350 Image Processing System. The 7350 consists of a color display monitor, a conversational monitor, and a control unit. The control unit is channel-attached to the host under VM support. Images are formed in a refresh buffer; the buffer entries are transformed by a color lookup table and sent to the color monitor. The color monitor is a 1024 × 1024 CRT with eight bits for each color (red, green, blue). The 7350 used in this study has six storage buffers; each is one megabyte. The refresh buffer is two megabytes. While transferring data between buffers, the Image Processor Arithmetic Logic Unit performs 16-bit arithmetic and Boolean operations on multiple-

buffer inputs. Also, buffer data can be transformed with lookup tables before or after processing and transfer. The 7350 has several window, mask, and selected-write features and a Random Address Processor that make it well suited to 3D imaging, animation, and multiple-window display. The display software is written in VS FORTRAN.

Host functions. As the frames are loaded into active memory from disk, several operations can be performed. The first and last frame numbers to be loaded are specified and a subimage can be selected. These two options can significantly reduce the storage used and the time required to form an image in the 7350. The time required to form a 3D image is directly proportional to the number of picture elements. The frames can be smoothed during loading. The resulting sequence of frames is saved in the 7350 storage buffer and is the base for subsequent imaging operations. The user interacts with the 3D data arrays in terms of physical dimensions, e.g., microns, volts, or miles. At any time during the image display operations the user can request a new set of variables from the host. Also, the multiple-window image formed in the refresh buffer can be saved on the host disk, or previously stored images can be restored.

Workstation functions. In the 7350 the 2D frames can be reviewed individually or in a rapid animated fashion. To form a 3D display on the color monitor, the user specifies the rotation of the input 3D array relative to the display plane. The display is then formed by projecting successive 2D frames of the data onto the display plane and by storing the result in the refresh buffer. The Random Address Processor in the 7350 control unit is used to map the input images onto the appropriate locations in the refresh buffer to form the specified rotation. Since all of the frames are resident in the 7350 storage buffers, 3D images can be formed in 5 to 15 seconds of real time. The 3D options do not increase the display time, since they are based on masking functions which are performed in parallel with the imaging.

Cited references

1. T. E. Graedel and R. McGill, "Graphical presentation of results from scientific computations," *Science* **215,** No. 4537, 1191–1198 (1982).
2. B. E. Brown, "Computer graphics for large scale two- and three-dimensional analysis of complex geometries," *Computer Graphics (Proceedings of SIGGRAPH 79)* **13,** No. 2, 33–40 (1979).
3. E. J. Farrell, S. E. Laux, P. L. Corson, and E. M. Buturla, "Animation and 3D color display of multiple-variable data: Application to semiconductor design," *IBM Journal of Research and Development* **29,** No. 3, 302–315 (1985).
4. C. G. Mohr and R. L. Vaughan, "Processing and display of multi-dimensional thunderstorm measurements," in J. J. Pearson (editor), *Processing and Display of Three-Dimensional Data II (Proceedings of SPIE)* **507,** Society of Photo-Optical Instrumentation Engineers, Bellingham, WA (1984), pp. 128–137.
5. E. J. Farrell, W. C. Yang, and R. A. Zappulla, "Animated 3D CT Imaging," *IEEE Computer Graphics and Applications* **5,** No. 12, 26–32 (1985).
6. C. M. Lubicki and K. W. Bedford, "Computer animation of storm surge predictions," *Journal of Hydraulic Engineering* **111,** No. 2, 284–299 (1983).
7. N. Ida and W. Lord, "Simulating electromagnetic NDT probe fields," *IEEE Computer Graphics and Applications* **3,** No. 3, 21–28 (1983).
8. E. J. Farrell, "Color display and interactive interpretation of three-dimensional data," *IBM Journal of Research and Development* **27,** No. 4, 356–366 (1983).
9. E. J. Farrell, R. Zappulla, and A. Kantrowitz, "Planning neurosurgery with interactive 3D computer imaging," *Proceedings of MEDINFO 86*, North-Holland Press, Amsterdam (1986), pp. 726–730.
10. E. R. Tufte, *The Visual Display of Quantitative Information*, Graphic Press, Cheshire, CT (1983).
11. J. N. Curtis and D. H. Schwieder, "Data analysis through a generalized interactive computer animation method (DATICAM)," *Computer Graphics* **9,** No. 2, 153–157 (1985).
12. J. L. Garbini, M. R. Lembersky, U. H. Chi, and M. T. Hehnen, "Merchandiser design using simulation with graphical animation," *Forest Products Journal* **34,** No. 4, 61–68 (1984).
13. D. E. Calkins and J. Ishimaru, "Computer graphics and animation come to ship design," *Computers in Mechanical Engineering* **3,** No. 1, 32–42 (July 1984).
14. G. H. Joblove and D. Greenberg, "Color spaces for computer graphics," *Computer Graphics (Proceedings of SIGGRAPH 78)* **12,** No. 3, 20–25 (1978).
15. G. W. Meyer and D. P. Greenberg, "Perceptual color spaces for computer graphics," *Computer Graphics (Proceedings of SIGGRAPH 80)* **4,** No. 2, 254–261 (1980).
16. S. M. Pizer, J. B. Zimmerman, and R. E. Johnston, "Contrast transmission in medical image display," *Proceedings of the IEEE Computer Society, ISMII '82* (1982), pp. 2–9.
17. A. Santisteban, "The perceptual color space of digital image display terminals," *IBM Journal of Research and Development* **27,** No. 2, 127–132 (1983).
18. J. P. J. deValk, W. J. M. Epping, and A. Heringa, "Colour representation of biomedical data," *Medical and Biological Engineering and Computing* **23,** No. 7, 343–351 (July 1985).
19. P. K. Robertson and F. O'Callaghan, "The generation of color sequences for univariate and bivariate mapping," *IEEE Computer Graphics and Applications* **6,** No. 2, 24–32 (1986).
20. J. R. Truckenbrod, "Effective use of color in computer graphics," *Computer Graphics (Proceedings of SIGGRAPH 81)* **15,** No. 3, 83–90 (1981).
21. J. Albers, *Introduction of Color*, Yale University Press, New Haven, CT (1963).
22. J. J. Pearson (editor), *Processing and Display of Three-Dimensional Data (Proceedings of SPIE)* **367,** Society of Photo-Optical Instrumentation Engineers, Bellingham, WA (1982).
23. J. J. Pearson (editor), *Processing and Display of Three-Dimensional Data II (Proceedings of SPIE)* **507,** Society of Photo-Optical Instrumentation Engineers, Bellingham, WA (1984).
24. D. L. Lauer, *Design Basics*, Holt, Rinehart and Winston, New York (1979).
25. L. Stroebel, H. Todd, and R. Zakia, *Visual Concepts for Photographers*, Focal Press Inc., New York (1980).

26. E. J. Farrell, R. A. Zappulla, and W. C. Yang, "Color 3D imaging of normal and pathologic intracranial structures," *IEEE Computer Graphics and Applications* **4,** No. 9, 5–17 (1984).

Edward J. Farrell *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598.* From 1959 to 1964 and from 1966 to 1968, Mr. Farrell was with UNIVAC. His work included reliability analysis of complex digital systems, waveform classification for sonar target identification, and spacecraft guidance system calibration and evaluation. He was at the Research Division of Control Data Corporation from 1964 to 1966. During that time, his research concerned the accuracy limits of scanning optical sensors for celestial navigation, and the design and performance of a star occultation scanner used for the navigation of the Gemini vehicle. In 1968, he joined the research staff at the Thomas J. Watson Research Center. His investigations have included the classification of patient recovery following open-heart surgery based on statistical and functional-simulation methods, medical ultrasonic tomography, and methods for color display of 3D data. His current interests are in the areas of interactive display and interpretation of multidimensional data with 3D color imaging, animation, multiple-window display methods, and menu/user interface design. Mr. Farrell received his B.S. from the University of Minnesota with a major in physics, and subsequently completed his doctoral course work in mathematical statistics.

Reprint Order No. G321-5293.