Color by Number: Imaging Large Data

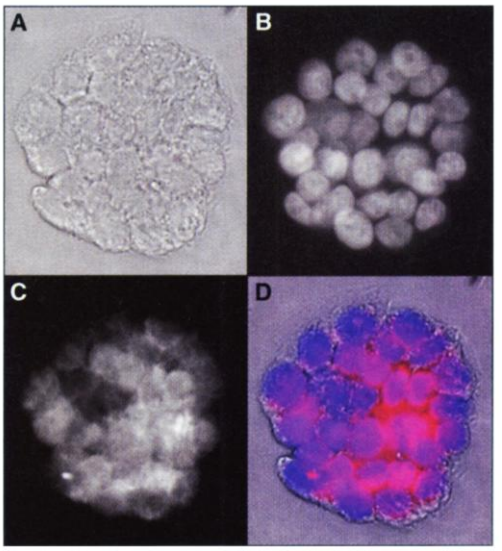
Carol A. Bertrand

Over the past few years, advances in acquisition hardware, storage capacity, and personal computer processing speed have made possible the ready acquisition and processing of enormous data sets. Large data sets, particularly multidimensional ones, can be difficult to analyze and present with standard numerical techniques. Images, however, can provide an effective means of displaying this information, especially when one uses color and brightness as tools to highlight statistically important features.

Computer images are visual representations of a matrix, a data set with *x*-*y* coordinates. Coordinate spacing of data may range from micrometers to miles, and values may represent any measurable parameter such as intensity of visible light or other electromagnetic radiation, temperature, color, sound, or force. When this matrix is displayed as an image, the values will be represented by pixels; therefore, a first step in image processing is defining how pixel brightness and color will best represent the data.

There are two standard image formats in common use, Indexed and Truecolor. The Truecolor, or RGB, format requires three numbers to define each pixel, corresponding to the amount of red, green, and blue that must be blended to achieve a specific color and brightness. The Indexed format uses only one number to define each pixel, and the depiction of that number in terms of color and brightness is established by palettes, color maps, or look-up tables (LUTs). The default palette in many popular programs is the grayscale, a linear distribution of grays between black and white. Standard image storage requirements are 8 bits per pixel for Indexed images and 24 bits per pixel (8 bits per color) for Truecolor.

In a typical grayscale image, pixels are assigned shades of gray from black to white (*y* axis) based on a straight line for values ranging from 0 to 255 (*x* axis). Programs used for image acquisition usually provide methods for visually enhancing an image display, as do general image processing software products, such as Adobe Photoshop (www.adobe.com). Common adjustments such as contrast and intensity manipulate the slope and intercept of this straight line, resulting in improved contrast between image regions or increased image brightness.



Co-localization study. Three separate images of an epithelial cell cluster were acquired under different conditions and then combined to assess the percentage of mucin-producing cells. (**A**) Transmitted light image of the cluster. (**B**) Hoechst-stained nuclei captured with ultraviolet fluorescence exposure. (**C**) Periodic acid Schiff-stained mucin captured with green fluorescence exposure. (**D**) Combination of the three images using the Truecolor format. Nuclei, blue channel; mucin, red channel. Images acquired and processed using SimplePCI software (www.cimaging.net).

If the intensity histogram indicates that the majority of values fall within a smaller region than 0 to 255, two types of adjustments can force the image to use the full range of grays. First, the scale can be made nonlinear by gamma adjustment, which changes the intensity gradients or steps in different regions of the scale. Alternatively, the value distribution can be stretched to occupy the full scale by assigning the minimum and maximum intensities of the image values of 0 and 255, respectively, and linearly reassigning intermediate intensities between the new endpoints.

If a data set's values consist of one measured parameter with 256 possible levels (8 bits of binary information), the Indexed format would be a first choice for display because it allows for a more informative display. Pseudocoloring (*1*) a data set that represents the surface temperature distribution of the Western hemisphere, for example, would allow easy color identification of the values. Shades of blue could be used for subfreezing temperatures, and shades of red, orange, and yellow could indicate higher temperatures. With pseudocoloring, the user creates a palette that specifically assigns the desired color to each corresponding level of information. Like the index of a book, the numbers stored in the data set identify the location in the palette that contains the desired pixel color and brightness.

An image captured with a digital color camera would, by default, contain the necessary color information for use of the Truecolor format. But Truecolor can be used in more applications than just simple display of standard pictures. The three colors used to define each pixel represent independent color channels, and each channel can be assigned a separate 8-bit data set. Because the combined levels of the independent channels define each pixel's color, the degree of correlation between data sets can be identified by color shifts. Assigning the previous data set of surface temperature distribution to the red channel and a data set of atmospheric CO_2 distribution to the green channel will produce an image that is red where temperature is high and CO_2 is low, green where CO_2 is high and temperature is low, and yellow when both gas and temperature are high. A similar channel assignment is used when performing a co-localization assay (*2*) in cell biology (see figure, this page). In this assay, used to determine whether two or more objects (e.g., proteins, organelles, structures) share similar regions, separate indexed images are acquired with the use of independent fluorescent labels. The images are assigned to separate color channels and then analyzed for color shifts to determine whether the labels occupy similar regions.

The objects or structures in an image can be analyzed for variations in size, position, shape, and intensity individually, as well as in relation to each other. Virtually all image processing programs provide an interactive environment for object identification that includes a visual display of the image, drawing tools that can be applied directly to the image, and a histogram or line graph showing the image in-

The author is in the Department of Cell Biology and Physiology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. E-mail: cbertra+@pitt.edu

tensity distributions. The aim of object identification is to separate the image into measurable regions (*1*), and the result is visually displayed as colored segments on a mask, a type of electronic stencil that stores the locations of the segments for recall during analysis or display.

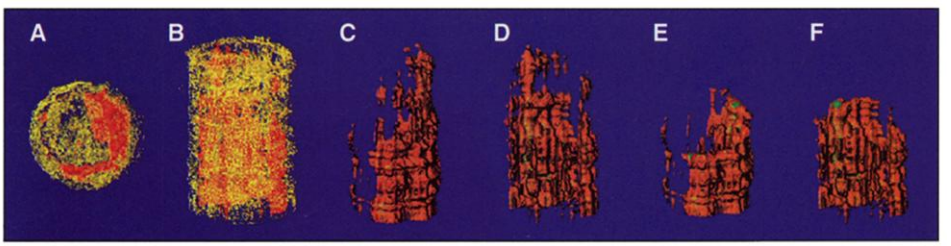
A standard technique for identifying objects that exhibit a narrow range of brightness or color is thresholding, which selects all pixels in an image that lie between a minimum and maximum threshold defined by the user. Ideally, when the objects to be identified are uniform and distinct from other features in the image, thresholding alone may be sufficient. In practice, the color or brightness distribution may not be uniform throughout the desired objects or structures, the transition from nonspecific background to object may be graded or faint, background noise can be present, and objects may touch or overlap and appear as aggregates after thresholding. Filtering operations will be required in these cases to assist the thresholding identification. Specific filters can smooth or normalize color and brightness variations in the interior of an object, clarify the edges of objects, remove nonspecific noise, or separate overlapping objects (*1*). Filtering does alter the data set, however, and should be performed on a copy of the image. The positions of objects identified after filtering and thresholding can be saved on a mask and applied to any image. Manual identification is also possible and involves drawing a boundary around each object, which is also saved on a mask.

After objects have been identified, numerical analyses can be performed on them. Operations include counting them, calculating their areas, identifying shapes, measuring distances between objects or user-defined reference points, and measuring intensity variations. Results are displayed in common spreadsheet formats that can be exported. Mathematical operators, including Boolean and standard operators (+, –, /, *, ^), can also be applied to images. These operators can be used to remove background noise by subtracting a test or calibration image from the desired image, or to normalize an image for multi-image comparisons.

The dimensions of an *x-y* data set remain fixed during an experiment. To observe changes over time, an individual plane can be resampled at suitable intervals to generate a series of images that can be viewed individually or sequentially as a movie. Uniform sampling intervals facilitate processing and display and, therefore, may be required in some of the more popular software packages. A three-dimensional (3D) image is created by incrementally scanning *x-y* planes up or down the *z* axis to collect a stack of images and then combining and displaying the stack with volume-rendering software (see figure, this page). Each individual plane in the stack must only contain data pertinent to that plane, unobscured by adjacent planes, for accurate 3D rendering. To accomplish this, the measurement apparatus must be able to focus on each plane separately during acquisition, as in confocal microscopy, or the data must be corrected after acquisition using mathematical processing such as deconvolution (*1*). It is also possible to sample a 3D object over time, referred to as 4D imaging (*3*).

The same types of analyses described for single images can also be performed on multiple images. To measure how a parameter changes during a time-lapse series, object identification is performed on the first image in the series and is saved to a mask. This mask is used with each image in the series, and the desired computations are repeated for each frame. This is the method used in cell biology to determine changes in cellular ion concentrations, such as calcium, over time. In the specific technique of ratiometric fluorescence microscopy, cells are identified in the first frame, and then the ratio of the mean fluorescent intensity at two specific wavelengths is used to calculate the cells' ion concentration in each frame of the series.

When an object moves during a time-lapse series, the process of object identification must be dynamic. In microscopy programs such as MetaMorph (www.universal-imaging.com), the user first manually identifies the particle to be tracked in the first frame of the series. The program then generates a box, centered on each particle to define its search region, steps to the next frame, locates the particle, and re-centers the box. The box size, which is set by the user, should be large enough to encompass the range of motion anticipated to occur in one time step but small enough to minimize the search time and avoid overlapping similar particles. After repeating the process for all images, the program will determine each object's velocity and pattern of movement. This technique may be used with a co-localization assay to demonstrate transient protein interactions.



Volume rendering of a stack. A scan of 50 *x-y* planes was performed on an epithelial cell after stimulated exocytosis and endocytosis in the presence of a fluorescent membrane label (*5*). The external label was washed out before scanning. The image is pseudocolored with yellow for low intensity, orange for medium, and green for high intensity. From the top (**A**) and full 3D view (**B**), the outer cell membranes appear yellow from faint labeling after washout but the endocytosed label appears orange. Lighting has not been applied. (**C**)The faint outer membrane label has been made transparent, and lighting and shading applied; the image is rotated to view the opposite side (**D**). Planar sectioning was used to remove the upper half of the image, revealing pockets of intense staining and absence of label in the region presumably occupied by the nucleus (**E** and **F**). Rendering was performed using VoxBlast software (www.vaytek.com).

Displaying 3D stacks as volumes requires specialized software to combine adjacent *x-y* data sets into volume data. After the planar data have been converted to their volume representation, the data set may be pseudocolored to highlight subset features; transparency and lighting can then be used to further distinguish objects (see figure, this page). Transparency allows one to remove backgrounds that obscure a desired object. A red fluorescent cell surrounded by a dark blue background would be rendered in three dimensions as a dark blue volume unless dark blue pixels are defined as transparent. Lighting gives a volume rendering a 3D appearance by applying shading and shadowing to objects on the basis of the source and direction of illumination. An additional feature available in volume rendering, and often used in medical imaging such as magnetic resonance imaging, is the use of planar sectioning of a volume, allowing an object to be visually dissected.

Imaging technologies have contributed to progress in virtually all areas of science, from the analysis of protein interactions in live cells to global atmospheric fluctuations (*4*). Digital imaging technology is at its best when it produces images that facilitate the interpretation and enhance the understanding of large, complicated data sets. The techniques of modern imaging will increasingly contribute to advances in science, thanks to improvements in acquisition technology and development of more powerful software tools.

References and Notes

1. J. C. Russ, *The Image Processing Handbook* (CRC Press, Boca Raton, FL, ed. 3, 1999).
2. A. Smallcombe, *Biotechniques* **30**, 1240 (2001).
3. S. Paddock, *Biotechniques* **30**, 283 (2001).
4. A. Lawler, *Science* **292**, 1044 (2001).
5. Images were collected at facilities funded by the Cystic Fibrosis Foundation.