Pictorial applications for range sensing cameras

V. Michael Bove, Jr.

Massachusetts Institute of Technology Media Laboratory
20 Ames Street, Cambridge, Massachusetts 02142

ABSTRACT

The acquisition of range information allows real-world scenes to be manipulated in ways traditionally restricted to computer-generated images. This paper will describe lighting, focus, and other changes upon real scenes, describe some of the implications of range sensing to efficient coding of image sequences, and outline some new research into passive range sensing cameras which use focus information to infer distance.

1. INTRODUCTION

Film, television, and photography, the three most ubiquitous representational visual media, typically present imagery to the viewer in a planar form. As the human visual system can correctly interpret these planar images as representing two-dimensional projections of three-dimensional reality from a specific viewpoint, it is not surprising that the camera itself performs the three-dimensional-to-two-dimensional projection. In sensing only a flat projection of reality, however, the camera discards a great deal of information that could be used to simplify various image-processing and understanding tasks.

In contrast with camera images, synthetic data bases as used in computer graphics are constructed in such a way as to enable rendering from variable viewpoints with variable lighting, segmentation of synthetic scenes and rearrangement of the constituent parts, and other variations. The intent of the research described in this paper is to develop a similarly rich and useful representation of real scenes and to explore the rendering variations and other operations made possible thereby. Using this new representation will redefine and expand the illustrative and creative ways in which images of real scenes are used.

2. TAKING PICTURES AND RANGE IMAGES

If it is desired to increase the flexibility of real images, a crucial piece of information besides the intensity value of each point in a scene (Figure 1) is the distance from the camera. When these distance values are arranged in register with the intensity image, the result may be displayed as a range image (Figure 2).

For industrial machine vision applications, rangefinding is commonly done using an active camera which projects a light source onto a scene and either triangulates or measures time-of-flight to determine distance.[1] The images in the preceding illustrations were taken with a fairly typical light-stripe laser rangefinder.

3. USING THE RANGE INFORMATION

In a lighting model commonly used for computer graphics, the intensity with which a particular point is rendered is a function of the normal vector to the surface at that point, the locations and intensities of all light sources incident at that point, and the reflectivity characteristics of the point, which are broken into specular and diffuse components:

$$I = I_{ambient} k_{ambient} + \frac{I_{source}}{r^2}[k_{diffuse}(\overline{\mathbf{L}} \cdot \overline{\mathbf{N}}) + k_{specular}(\overline{\mathbf{R}} \cdot \overline{\mathbf{V}})^n] \quad (1)$$

where $\overline{\mathbf{L}} \cdot \overline{\mathbf{N}}$ is the cosine of the angle between the unit surface normal and the unit vector to the light source, and $\overline{\mathbf{R}} \cdot \overline{\mathbf{V}}$ is the cosine of the angle between the unit vector in the direction of reflection and the unit vector toward the viewpoint.[2]

This formula is in fact an empirical approximation rather than a physical model, but realistic results can be obtained with appropriately selected values for the proportionality constants k and for the exponent n (Figures 3,4). (For a color image, this model must be applied three times to the red, green, and blue components of the image, and for multiple light sources the contributions from each source must be summed at each point.)

Producing an array of surface normals $\overline{\mathrm{N}}$ for a real scene requires differentiating the range image across x and y. In order to permit the lighting to be manipulated computationally, the intensity image should be free of strong shadows or highlights, and should ideally represent the reflectivity of each point in the scene. In the case of the illustrations, the scene was simply illuminated with very diffuse lighting. Alternatively, it should be possible to undo the effects of ambient lighting since surface orientation is known. As the specularity or diffuseness of a point is not currently sensed by the camera system, all points are treated as diffuse reflectors. For a more realistic rendering, the camera should be made capable of measuring reflectance characteristics, or several different sets of characteristics (*e.g.* purely diffuse, glossy, purely specular) should be pre-programmed into the system and the lighting software assign different ones to different known objects.

Another operation possible upon a scene for which range is known is projection from variable viewpoints (Figure 5). As the camera cannot see through objects, the background will contain visible "holes" when the viewer looks around foreground objects. Methods are being investigated to minimize this effect by extrapolating surrounding surfaces into these holes, combining the output of more than one camera, or (in the case of a movie) using background information seen at some other time.

If the reflectivity image of a scene is taken with a small lens aperture, all points will be in focus, and it becomes possible to simulate the effect of some other lens with a shorter depth of field. When a point source of light is imaged by a lens and aperture, its energy is spread over a region whose area increases as the point moves away from the focus distance of the lens. Calculating a defocused image requires modeling each pixel in the in-focus scene as a point source of light and solving the Huygens-Fresnel integral for the case of Fraunhofer diffraction of a point source of light by a circular aperture. The integral cannot be evaluated in closed form, but its value may be calculated as a series of Lommel functions, which are themselves infinite sums of Bessel functions.[3]

The resulting set of equations describes the distribution only for a single wavelength of light. In white light, the maxima of one frequency may fall upon the minima of another, with the result that the distribution tends toward a uniform circle, particularly in imaging systems where the resolution is coarse and aperture large compared with wavelength. Photographers use this uniform approximation, calling the defocused spot the "circle of confusion."[4]

Deriving an expression for the diameter of this circle is a straightforward geometric problem. The lens equation says that the reciprocal of the object-to-lens distance and the reciprocal of the lens-to-focus distance sum to a constant, which is the reciprocal of the focal length of the lens:

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}. \tag{2}$$

In Figure 6, a camera lens is focused such that a point X produces a sharp point image on the film or image sensor at F. Point Y, which is farther away, focuses in front of the focal plane at point C; on the film the light energy from that point is distributed over a larger region of diameter c. Since triangles ABC and DEC are similar,

$$c=d\frac{V_x-V_y}{V_y}. \tag{3}$$

But lens diameter d can also be expressed in terms of the focal length f and the numerical aperture (or f-number) n:

$$d=\frac{f}{n} \tag{4}$$

and since a similar geometrical construction is valid for points Y closer than X (which will focus *behind* the focal plane),

$$c=\frac{f}{nV_y}|V_y-V_x|. \tag{5}$$

The circle-of-confusion function is shown in Figure 7. This function is quite asymmetrical about the plane of best focus. As point Y moves toward infinity, the diameter approaches a limit $(V_x/n - f/n)$, while it increases much more rapidly

for points closer to the lens than point X. At $U_y = f$ the circle of confusion is the same as the effective diameter of the stopped-down lens (here the lens is behaving as a collimator). The result of applying simulated defocusing to a real scene is shown in Figure 8.

Depending on the computer-graphics methods one prefers to use, the range information may be used to construct a particle or polygon representation for a scene, with the reflectivity image providing a texture map for the surfaces or particles. Information about the simulated light sources and desired viewpoint, as well as the focal length, *f*-number, and focus distance of a simulated lens may then be used to project, illuminate, and defocus the scene.

Any operation that can be performed on a still image can likewise be performed on the frames of a movie, permitting interesting and useful cinematographic variations to be made at the editing rather than shooting stage. But the impact of sensing range information for a moving image is potentially much greater than that. As the databases used in computer animation are far more compact to store and transmit than the movies that are generated from them, it would be desirable to go the opposite direction and represent real moving scenes as a set of objects defined once and given three-dimensional trajectories. Not only would such a representation offer coding efficiency, it could also eliminate the concept of the "frame" altogether, effectively decoupling the spatial and temporal resolution of the camera from that of the final presentation.

4. A PASSIVE RANGE CAMERA

While an active camera system such as a laser rangefinder can give excellent accuracy, it is restricted to relatively small indoor scenes and (unless very specialized hardware is used) is too slow for capturing range information for moving objects. This consideration has led to a search for rangefinding techniques which inherently offer greater speed and which do not require contrived lighting.

The aforementioned simulation of a lens and aperture model suggests the possibility of carrying out the process in reverse, namely using focusing to derive the distance to points in a scene. The defocusing process may be defined as the convolution of image intensities $i(x, y)$ with a point-spread function $f(z)$ which is a function of distance to each point:[5]

$$i_{defocused}(x, y) = i(x, y) * f(z(x, y)). \tag{6}$$

The goal of depth-from-focus is to recover $f(z)$, and knowing its form, to calculate z.[6] The recovery becomes practical if carried out in the frequency domain, where convolution transforms to multiplication. Taking the Fourier transform of (6) and rearranging,

$$F(\omega_x, \omega_y) = I_{defocused}(\omega_x, \omega_y)/I(\omega_x, \omega_y). \tag{7}$$

Since the point-spread function is circularly symmetrical, its transform should be circularly symmetrical as well, and the problem may be solved in one dimension (that of radial distance) rather than two.

In practice, the camera is focused slightly in front of the closest object in a scene, and two images are produced: one with a small aperture (say *f*32) and one with a larger aperture (Figures 9,10). The optimal aperture difference will depend on the overall depth of the scene being measured: if the two apertures are close in size, it will be quite some distance before measurable defocusing takes place, while a large difference will produce greater accuracy in the foreground, but error in the background where the circle-of-confusion function approaches its asymptote (recall Figure 7). A neutral density filter is used to reduce the light coming through the lens when its aperture is opened up.

The assumption is made that the long-depth-of-field image approximates perfect focus at all distances, while the short-depth-of-field is the result of convolving these perfectly focused image points with a distance-dependent point-spread function. A two-dimensional discrete Fourier transform (DFT) is taken of corresponding windowed blocks from the two images, and the resulting two-dimensional power spectra are collapsed down to one dimension by averaging all power at given Cartesian distances from $(\omega_x, \omega_y) = (0, 0)$. The short-depth-of-field spectrum is then divided by the long-depth-of-field spectrum, and the resulting values (which represent the power spectrum for the point spread function) are used to solve a regression equation for the diameter of the circle of confusion at each image point; this is then converted to distance (Figure 11). A special two-iris lens is being built which will create side-by-side long- and short-depth-of-field images with a film or video camera, allowing this process to be extended to moving scenes.

Depth-from-focus requires sufficient high-frequency content in the image that perceptible defocusing takes place, and cannot estimate the distance to completely uniform regions. Also, in order to assure that there are enough power

spectrum terms to allow good estimation, fairly large regions (32-by-32 or 64-by-64 pixel blocks) must be transformed, with corresponding loss of computational speed and localization of depth values. The shape of the circle-of-confusion curve itself suggests further problems. Clearly no range resolution is possible until the point-spread function is significantly larger than the smallest detail resolvable by the camera, while for very distant points the curve comes close to horizontal and large changes in distance produce imperceptible variations in the diameter. Noise in the input images also affects the accuracy of this method.

While clearly not as good as structured-light rangefinding, depth-from-focus has shown capable of producing three to four bits of useful range information, and if combined with other passive camera methods of range sensing may be able to provide much improved accuracy.[7] Even at the current resolution, depth-from-focus data prove useful for image segmentation and keying, where objects from one scene are placed onto another background (Figures 12,13)

Additionally, research elsewhere has demonstrated that even coarser range information than that provided by this implementation of the depth-from-focus process may be used to create parameterized shape descriptions of real scenes, and to identify objects in a scene from a catalog of known objects.[8]

5. CONCLUSIONS

The research described above aims toward a new way of thinking about photography and moviemaking, where the camera creates not a series of frames but a compact, modifiable, and variably renderable database akin to those used in computer graphics. Simply by sensing range information, the camera enables a set of interesting and useful variations to be performed upon the images taken. Methods are being developed by which the taking apparatus need not be significantly different to the user from ordinary cameras.

6. ACKNOWLEDGMENTS

The author wishes to thank Andrew Lippman and Alex Pentland of the MIT Media Laboratory for their assistance and helpful suggestions, and Hubert Delany for the variable viewpoint rendering software. The research work described in this paper has been supported by CPW Technologies and International Business Machines.

7. REFERENCES

1. R. A. Jarvis, "A Perspective on Range Finding Techniques for Computer Vision," IEEE PAMI-5(2), pp. 122-139 (1983).

2. J. D. Foley and A. van Dam, Fundamentals of Interactive Computer Graphics, pp. 575-578, Addison-Wesley, Reading, Massachusetts (1982).

3. M. Born and E. Wolf, Principles of Optics, pp. 435-443, Pergamon Press, Oxford, England (1970).

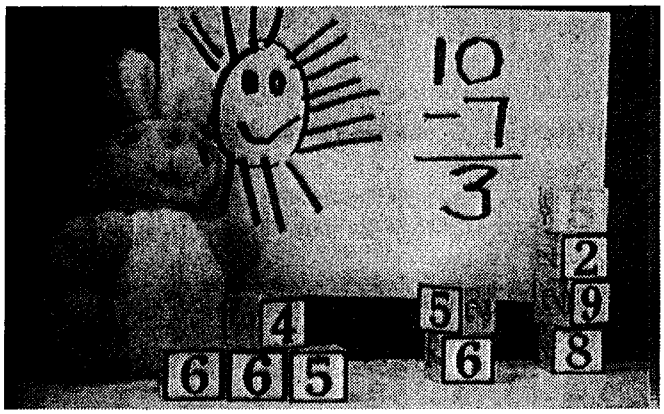
4. P. E. Boucher, Fundamentals of Photography, pp. 53-61, Morgan and Morgan, New York, New York (1968).

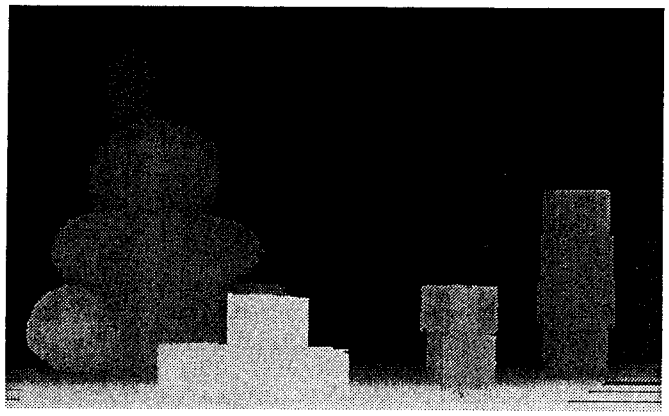
5. B. Horn, "Focusing," Project MAC Artificial Intelligence Memo No. 160, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, Massachusetts (1968).

6. A. Pentland, "A New Sense for Depth of Field," IEEE PAMI-9(4), pp. 523-531 (1987).

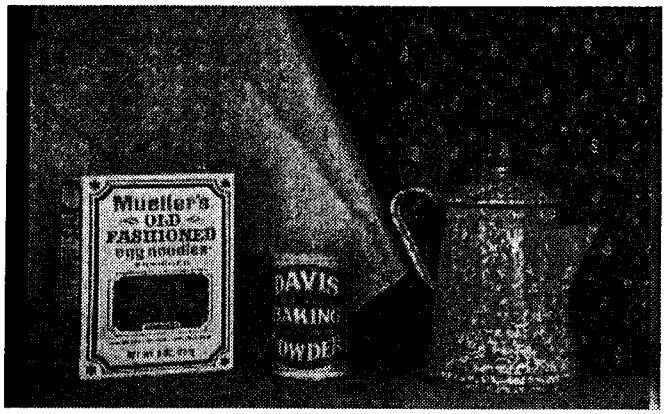
7. Jarvis, *op. cit.*

8. A. Pentland, "The Parts of Perception," Report No. CSLI-87-77, Center for the Study of Language and Information, Stanford, California (1987).





Figures 1, 2: Image of a scene, and its range (lighter objects are closer).



Figures 3, 4: Knowing the range information makes it possible to use a computer to vary the lighting on a real scene.

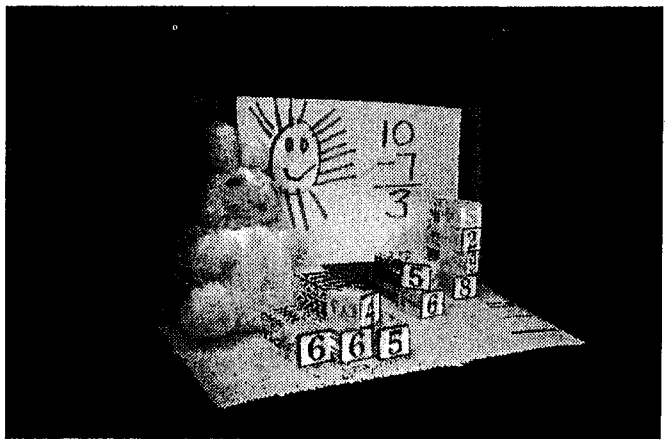


Figure 5: The image of Figure 1 projected from a different viewpoint.

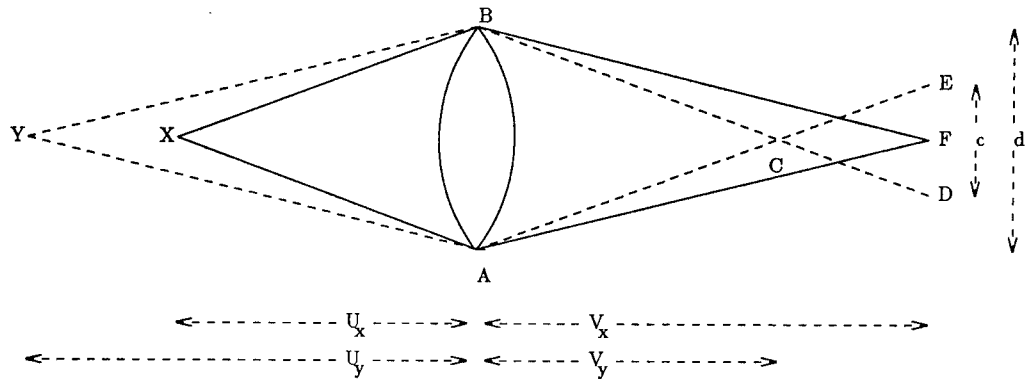


Figure 6: An in-focus point X images as a point at distance V_x. Out-of-focus point Y images as a circle of diameter c.

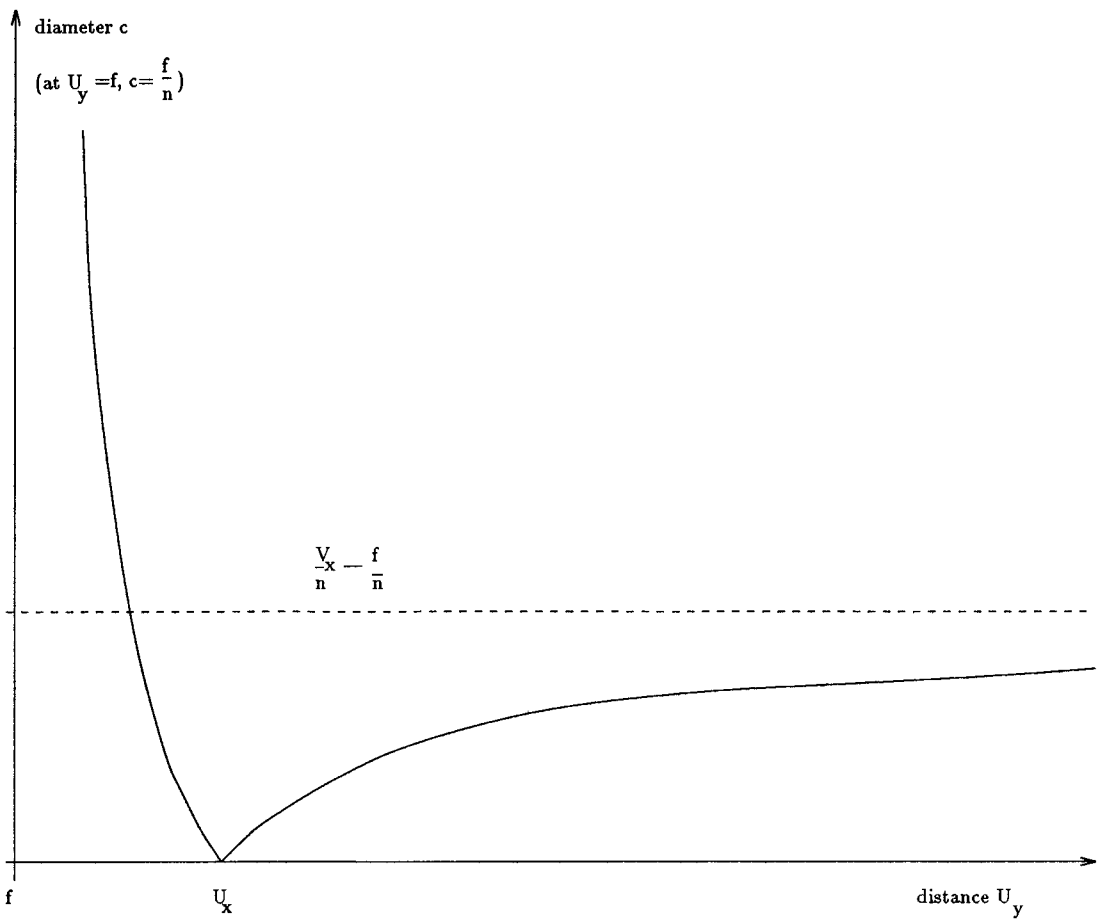


Figure 7: Illustration of the circle-of-confusion function.

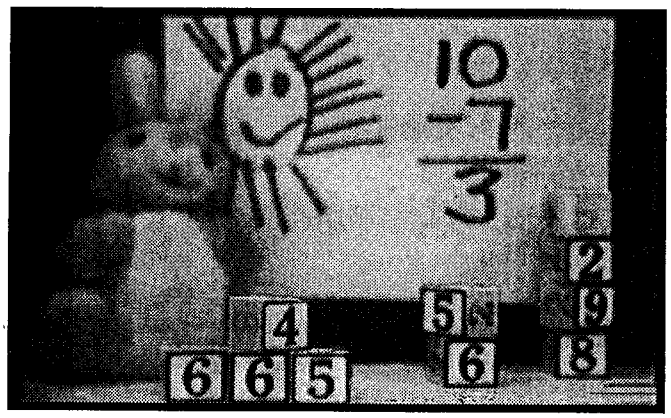
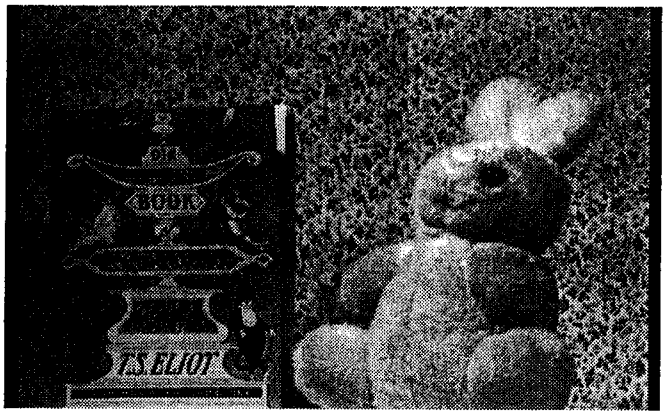
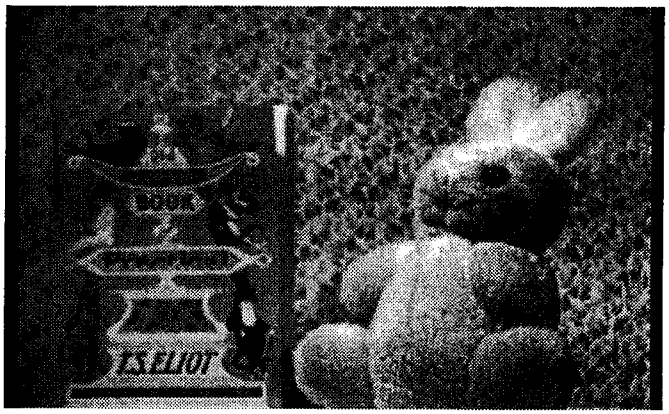


Figure 8: The image of Figure 1 with a simulated short-depth-of-field lens focused on the front row of blocks.





Figures 9, 10: Calculating depth-from-focus requires taking two pictures, one with a small lens aperture (left) and one with a larger aperture (right).

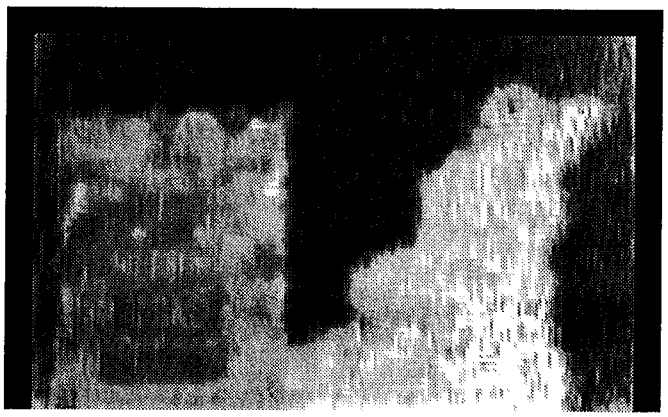
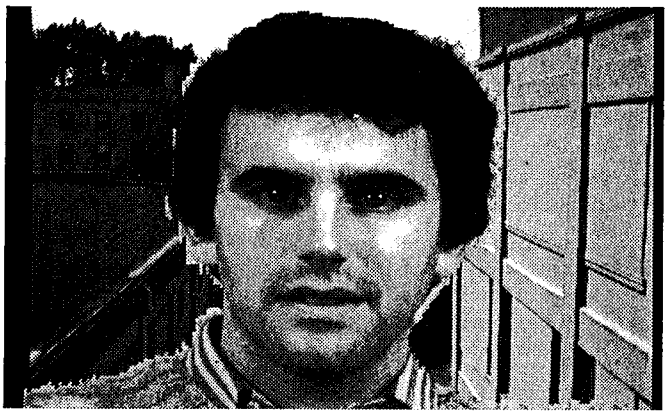


Figure 11: Depth-from-focus data calculated from the preceding two images.



Figures 12, 13: Depth-from-focus data can be used to create a key mask (left) for placing a figure against a different background (right).