Single Lens Stereo with a Plenoptic Camera

Edward H. Adelson and John Y.A. Wang

Abstract—Ordinary cameras gather light across the area of their lens aperture, and the light striking a given subregion of the aperture is structured somewhat differently than the light striking an adjacent subregion. By analyzing this optical structure, one can infer the depths of objects in the scene, i.e., one can achieve "single lens stereo." We describe a novel camera for performing this analysis. It incorporates a single main lens along with a lenticular array placed at the sensor plane. The resulting "plenoptic camera" provides information about how the scene would look when viewed from a continuum of possible viewpoints bounded by the main lens aperture. Deriving depth information is simpler than in a binocular stereo system because the correspondence problem is minimized. The camera extracts information about both horizontal and vertical parallax, which improves the reliability of the depth estimates.

I. INTRODUCTION

"**EVERY BODY** in the light and shade fills the surrounding air with infinite images of itself; and these, by infinite pyramids diffused in the air, represent this body throughout space and on every side." Leonardo da Vinci [1] uses these words, together with the drawing in Fig. 1, to describe the relationship between objects' light, and image formation. The object in the drawing sends off rays of light in all directions, and if we choose to place a pinhole camera at any given point in space, we will discover that an image is formed. The image is the projection of a cone of light that Leonardo called a "visual pyramid." The space surrounding an object is densely filled with these pyramids, each representing an image of the object from a slightly different point of view. These infinitely multiplexed images, Leonardo emphasized, are simultaneously present throughout space whether we are aware of them or not.

If an ordinary pinhole camera is placed near an object, it selects a single visual pyramid and forms a single image, as shown in Fig. 2(a). If a pair of adjacent pinhole cameras are used, then two different images are formed, as in Fig. 2(b). The two images together give additional information about the structure of the light surrounding the object, and they help constrain the interpretation of the object's 3-D form.

Binocular stereo systems [2] extract depth information by the use of two cameras, which are normally treated as pinhole cameras. Binocular stereo is perhaps the most popular method of passive depth measurement. It can be highly effective in some situations, but it is known to suffer from some problems.

Two separate cameras muxt be used, which increases the bulk and expense of the system and leads to difficulties with camera calibration. There are ambiguities ahout correspondence that must be solved which can present formitiable computational challenges. In addition, binocular stereo exploits parallax along one axis and cannot offer depth estimates for contours parallel to this axis.

Manuscript received October 20, 1990; reviewed January 2 1991. This work was supported by a contract with SECOM Inc., Japan.

The authors are with the Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02134.

IEEE Log Number 9102680.

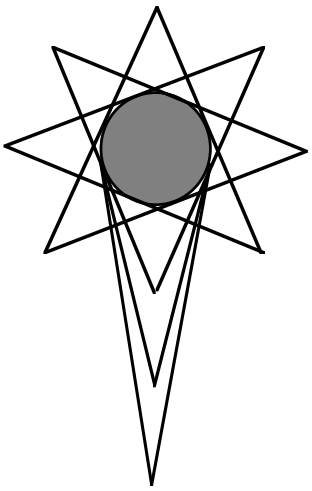


Fig. 1. Diagram from Leonardo's notebooks illustrating the fact that the light rays leaving an object's surface may be considered to form a collection of cones (which Leonardo calls "pyramids"), each cone constituting an image that would be seen by a pinhole camera at a given location.

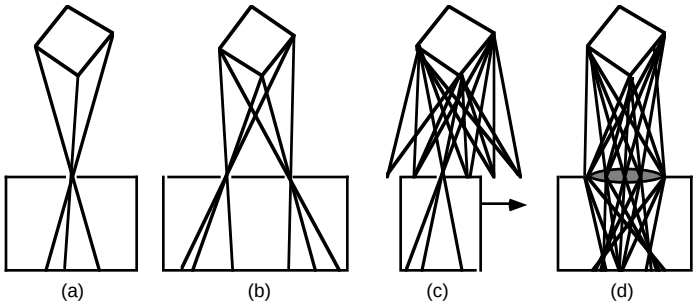


Fig. 2. (a) Pinhole camera forms an image from a single viewpoint; (b) in a stereo system, two images are formed from different viewpoints; (c) in a motion parallax system, a sequence of images are captured from many adjacent viewpoints; (d) a lens gathers light from a continuum of viewpoints; in an ordinary camera these images are averaged at the sensor plane.

The latter two problems can be ameliorated to some extent by the use of a trinocular system [3], [4], but this involves a third camera and the attendant increase in size, expense, and calibration.

Rather than taking just two pictures one can move a camera along a track and take a dense sequence of images, as shown in Fig. 2(c). Such a system can extract depth through the use of motion parallax [5].

A camera with a lens, such as that shown in Fig 2(d), gathers light from a continuum of viewpoints, i.e., it takes in the con-

tinuum of Leonardo's pyramids that have their apexes in the lens aperture plane. In an ordinary camera, the final image is simply the average of all the individual images that would be seen through all of the different positions of the lens aperture. These images will all be somewhat different for objects off the plane of focus, and objects become blurred according to their distance from the plane of focus. Pentland [6] has described a method for taking advantage of the properties of light within the lens aperture by comparing two images taken with two apertures of different sizes; the change in blur can be used to estimate the depth (see also [7]).

It is clear from the illustration in Fig. 2(d) that the structure of light passing through a lens is quite rich and potentially informative. Unfortunately, most of the structure is lost at the final moment of image formation when all of the light rays are projected onto a single planar surface and each sensor element registers the average of all the light rays striking it from different angles

The full complexity of the optical information filling space can be formalized as a single function, which Adelson and Bergen [8] call the "plenoptic function," where *plenoptic* is derived from the word roots for "complete" and "view." The plenoptic function describes the intensity of each light ray in the world as a function of visual angle, wavelength, time, and viewing position. It captures everything that can potentially be seen by an optical device and is related to what Gibson [9] called the structure of ambient light.

We will describe a camera that captures a chunk of the optical structure of the light impinging on the lens. The camera records information about how the world appears from all possible viewpoints within the lens aperture, and we refer to it as a "plenoptic camera."

Our camera gathers light with a single lens but uses a lenticular array at the image plane to retain information about the structure of the light impinging on different subregions of the lens aperture. After this information has been recorded, one can select the image information that came from any subregion of the aperture. Thus, one can simulate the image that would have been seen if the viewing position had heen shifted slightly up, down, left, or right. Moreover, one can measure the parallax corresponding to these virtual displacements and thereby derive depth estimates for objects in the scene. The extent of the virtual displacement and, therefore, the potential resolution of depth, is limited by the size of the lens aperture and tends to be lower than that for binocular stereo systems. Nonetheless, the technique offers some distinct advantages. Only a single camera is required; the correspondence problem is minimized, and parallax is available in both the horizontal and vertical directions.

II. SINGLE LENS STEREO

Before describing the plenoptic camera, let us explore the optical information available across the aperture of a camera lens. Suppose the camera is bringing a point object into focus on its sensor plane, as shown in Fig 3(a). If the object is placed nearer or farther from the camera as in Fig 3(b) and (c), the image is thrown out of focus. The luminance profiles of three images are

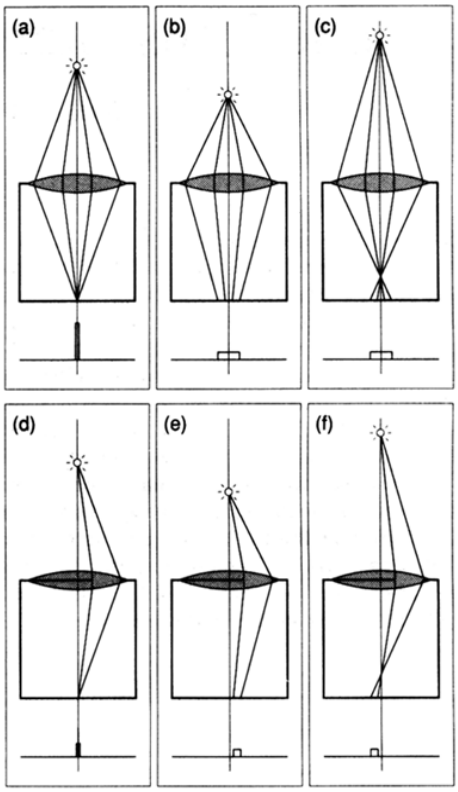


Fig. 3. Principle of single lens stereo: (a) In-focus point object forms a point image; (b) near object; (c) far object forms a blurred image; (d) with an eccentric aperture, the image of the in-focus object retains its position, but the images of the near or far objects (e) and (f), are displaced to the right or left.

shown just beneath the figures of the cameras the in-focus image is a bright small point, whereas the out-of-focus images are broader and dimmer. If the lens aperture is a circular disc, then the point spread function (PSF) will be a pill box, where the diameter of the pill box is a function of the aperture size and the degree of defocus. In one dimension, as shown here, the PSF of a hard-edged aperture will be rectangular (neglecting the minor effect of diffraction).

Now consider what occurs if we place an eccentric aperture in the lens. If the object is in the plane of focus, as in Fig. 3(d), then its image remains sharp and at the same position, but if the object is near, as in Fig. 3(e), then its (blurred) image is displaced to the right of center because the aperture selectively passes those rays that fall to the right-hand side. Conversely, if the object is far, as in Fig. 3(f), then its image is displaced to the left. The extent of displacement is proportional to the amount of blurring resulting from the defocus. (This fact turns out to be significant in reducing the correspondence problem, as will be discussed later).

As the aperture is displaced to the left or right, the image of a near object is likewise displaced to the left or right, whereas the image of a far object is displaced to the right or left. The amplitude and direction of the displacement allows one to determine the object's distance.

One can therefore acquire a sequence of images in which the aperture is displaced and can perform displacement analysis on the sequence in order to determine the depths of the objects in the scene. This method can be termed "single-lens-stereo."

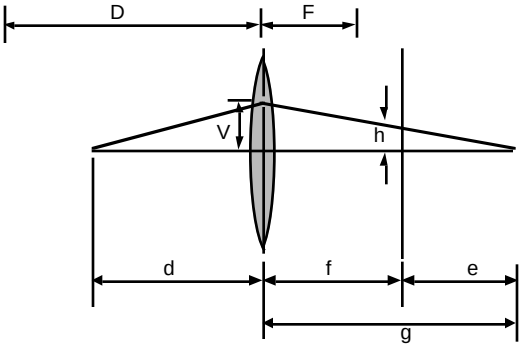


Fig. 4. Geometry of single lens stereo.

III. GEOMETRICAL OPTICS

The relationship between displacement and depth in a single-lens-stereo system may be understood by referring to Fig. 4. Consider a lens that is forming an out-of-focus image of a point object, and suppose that the lens is equipped with an eccentric point aperture. Let

- F focal length of lens
- f distance between lens and sensor plane
- D distance to a plane conjugate to sensor plane
- d distance of a particular point object
- g distance to conjugate focus of object
- e distance of conjugate focal point beyond sensor plane
- v displacement of aperture
- h displacement of object's image in sensor plane.

We would like to determine the object distance, d, given the known aperture displacement v and the resultant image displacement h.

By the use of similar triangles

$$\frac{1}{g} = \frac{1}{f}\left(1 - \frac{h}{v}\right) \tag{1}$$

and by the lens equation

$$\frac{1}{F} = \frac{1}{g} + \frac{1}{d} \tag{2}$$

which leads to

$$\frac{1}{d} = \frac{1}{F} - \frac{1}{f}\left(1 - \frac{h}{v}\right) \tag{3}$$

or

$$\frac{1}{d} = \frac{h}{v}\left(\frac{1}{F} - \frac{1}{D}\right) + \frac{1}{D} \tag{4}$$

The left-hand side is the reciprocal of the object distance d; the right-hand side consists of the known system parameters F, D, and v, along with the measured image displacement h.

IV. THE PLENOPTIC CAMERA

The single-lens-stereo approach will work as described but it has the disadvantage of requiring the accumulation of several snapshots over time. One would prefer to acquire all of the image information with a single snapshot. This demands a special optical system, which we will now describe.

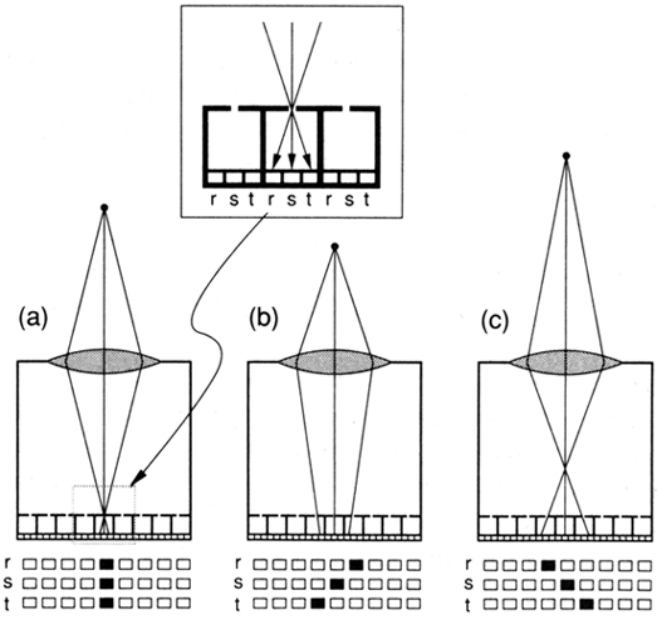


Fig. 5. Array of miniature pinhole cameras placed at the image plane can be used to analyze the structure of the light striking each macropixel.

In an ordinary camera, all of the light striking a given photodetector element (e.g., one cell of a CCD array) is treated in the same way: The photon responses are summed, regardless of the angle of incidence. If we could somehow keep track of the amount of light striking the photodetector from different directions then we could determine how much light came from the various subregions of the lens aperture.

Consider the arrangement shown in Fig. 5. The sensor array is covered with an array of tiny pinhole cameras. The light impinging on each point of the image is broken up into three subparts, each corresponding to a particular angle of incidence. An example for the case of an object in the plane of focus is shown in Fig. 5(a); the inset shows an enlargement of the pinhole array system. The image may be considered to be formed of macropixels, corresponding to the individual pinhole cameras, and each macropixel is subdivided into a set of three subpixels. The subpixels are of three types here and are labelled r, s, and t. Light passing through the right side, center, or left side of the lens will strike the r, s, or t pixels, respectively.

In effect, each tiny pinhole camera forms an image of the main lens aperture, and this image captures the information about which subset of the light passed through a given subregion of the main lens. (In order for this to work correctly, the pinhole cameras must be aimed at the center of the main lens; in Fig. 5, this is accomplished by displacing the pinholes as a function of eccentricity. An alternate approach is to use a field lens as discussed later.)

If the object is in the plane of focus, as in Fig. 5(a), then all three of the pixels r, s, and t of the center macropixel are illuminated. If the object is near or far, as in Fig. 5(b) and (c), then the light is distributed across the pixels in a manner that is diagnostic of depth. A good way to characterize this distribution is to create separate subimages from the r, s, and t pixel groups. The r subimage corresponds to light passing through the right side of the main lens, the s subimage to light passing through the center, and the t subimage to light passing through the left.

The three subimages, which are labeled *r*, *s*, and *t*, are shown beneath cases (a) (b) and (c) of Fig. 5. When the object lies in the plane of focus, as in (a), the three images are aligned. When the object is near, as in (b), the images are displaced successively to the left. When the object is far, as in (c), the images are displaced successively to the right. By measuring the displacement, one can estimate an object's depth.

The pinhole array can he replaced by an array of microlenses, i.e., a lenticular array. The lenticular array may consist of 1-D (cylindrical) lenses or may consist of a set of 2-D (spherical) lenses; the latter is sometimes called a fly's eye array. Lenticular systems can offer improved light-gathering efficiency and reduced aliasing artifacts since the entrance pupil can be much larger than with a pinhole system. We may think of the microlens array as constituting a battery of tiny cameras, where each tiny camera forms an image of the main lens aperture as seen from a different position in the sensor plane.

Students of photographic history will recognize that the camera involves the optical principles laid out by Lippman [10] and Ives [11] in their experiments on 3-D photography early in this century; the same principles are sometimes used to produce 3-D displays [12] and autofocus systems of some 35 mm SLR cameras [13].

A more fully developed optical system is shown in Fig. 6(a). The object at the left is imaged by the main lens onto a lenticular array, behind which lies a sensor array such as a CCD. Each lenticule gathers light over one macropixel and forms a microimage of the main lens aperture on the sensor elements beneath it. The field lens places the main lens aperture at optical infinity from the lenticules so that rays from the center of the main lens are imaged in the center of each microimage. A weak diffuser may also be used (shown here in front of the main lens but really placed in the main lens aperture plane) to prevent aliasing due to the sampling of the image by the lenticules. The weak diffuser is designed to blur the image very slightly with a point spread function roughly as wide as the spacing between lenticules.

Rather than imaging directly onto a sensor array, one can use a relay lens to form the image on a separate sensor array, i.e., one can view the aerial image with a video camera. This relay system, which is illustrated in Fig. 6(h), is more cumbersome than the direct imaging system, but it has the advantage of flexibility. We have used a relay system in our experiments because it allows us to match the effective size of the aerial image with the size of our CCD sensor. The relay system may introduce difficulties with vignetting, which can be improved by placing a fine groundglass diffuser in the image plane at a cost of reduced light intensity.

V. DESIGN CONSIDERATIONS

If each macropixel is divided into n subpixels along one dimension, then the spatial resolution of the system is reduced by a factor of n in that dimension. Thus, if the macropixels are 5×5, a system with a 500×500 element CCD will offer a spatial resolution of only 100×100. In principle, one could obtain useful depth information from macropixels as small as

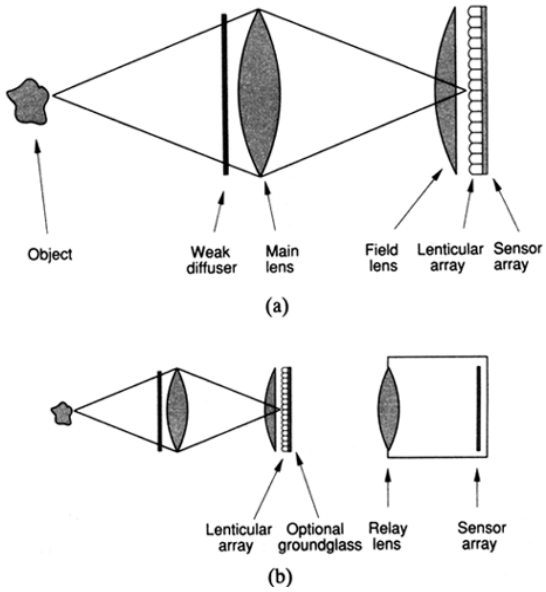


Fig. 6. (a) Optical system of a plenoptic camera; (h) plenoptic camera utilizing relay optics.

2×2, but in our experiments, we have typically used sizes on the order of 5×5. In our prototype system, the pixels at the very edges of the macropixels are often unusable due to vignetting and light spillage from adjacent macropixels; these difficulties might be overcome with a custom fabricated sensor.

Spatial resolution and parallax resolution trade off against one another. If a macropixel is divided into n subpixels, one can derive n views of slightly different positions by addressing different subpixels. More important, the depth-of-field of each image increases in proportion to n. Since the plenoptic camera begins with a large-aperture lens, objects out of the plane of focus become severely blurred, which limits the spatial resolution for these objects. This blurring removes fine detail and texture, which could otherwise be an important source of displacement information. The effective aperture size for a subimage is proportional to $1/n$ times the original aperture size. The best performance would be obtained with a very high-resolution sensor; it would be advantageous to use a sensor with far higher resolution than the standard video CCD used in our experiments.

If the lenticular array is cylindrical (i.e., 1-D), then the loss of spatial resolution only occurs in one dimension, but the parallax information is only acquired in one dimension.

The effective stereo baseline is limited by the physical size of the main lens aperture. Therefore, it is advantageous to use a main lens with a large aperture, i.e., a small *f*/number. However, a large aperture also leads to a shallow depth of field, which may be disadvantageous in some applications. The depth of field in each subimage is determined by the effective aperture size, and therefore, a greater depth of field can be attained by increasing the number of subpixels that are used to divide each macropixel. In a lenticular system, the *f*/number of the main lens should be matched to the *f*/number of the lenticular array.

Depth resolution depends on the ratio of the lens aperture to the object distance. Therefore, the best depth resolution is

obtained for small objects that are relatively close to the camera. For this reason, single-lens stereo systems may be better suited to close-range applications such as the inspection of parts rather than distant-range applications such as vehicle navigation.

VI. IMAGE PROCESSING

After a plenoptic camera's image is digitized, the array of macropixels may be analyzed to obtain a set of subimages. The subimages are extracted by selecting the appropriate subpixels of each macropixel by applying a weighting mask shifted to the various desired positions. The next problem is to perform displacement analysis on the set of images so obtained. The most straightforward approach is to apply standard two-frame displacement analysis to the image pairs obtained from adjacent subpixels. Each image pair gives a displacement estimate, and these multiple estimates can be combined. For example, if we begin with a 3×3 array of subimages, then there are a total of 12 image pairs that can be used to estimate displacement: six in the horizontal direction and six in the vertical direction.

The displacement analysis is simpler for the plenoptic camera than it is in the usual binocular stereo situation because the correspondence problem can be minimized or avoided. The classical correspondence problem occurs in motion or stereo when the displacement between two, images is greater than half of a wavelength of the highest frequency component in the image. An image that has been spatially low-pass filtered (i.e., blurred) below a cut-off frequency of ω can endure a displacement of up to $1/(2\omega)$ before the correspondence problem sets in. Thus, the correspondence problem is simply a manifestation of aliasing; it arises when the two images being compared have been sampled without sufficient prefiltering. In the case of motion analysis, the aliasing is in the time dimension, and it can be reused by averaging over time in order to generate motion blur. In the case of binocular stereo. the aliasing is in the dimension of viewing position (e.g., v_x or v_y); it can be reduced in principle by averaging over viewing position, i.e., through optical blur, but would only be effective if the apertures of the two cameras were so large as to be optically overlapping, which is not generally the case.

In the case of the plenoptic camera, it is possible for the virtual aperture of one subimage to abut or even overlap the virtual aperture of the next; insofar as these virtual apertures provide sufficient low-pass filtering aliasing can be prevented. Moreover, the optics automatically scale the prefiltering by an appropriate amount, as may be seen by consideration of the optics in Fig. 3. As an object moves away from the plane of focus, its parallactic displacement grows linearly, but its optical blur also grows linearly. Insofar as the optical blur is sufficient to prevent aliasing for one depth, it is automatically sufficient to prevent it for any depth. For this reason, correspondence ambiguities are minimized or avoided.

In practice, we have not experienced significant difficulties with correspondence, and we are therefore able to use a simple one-pass displacement algorithm without resorting to any procedures involving search, coarse-to-fine processing, or iteration.

There is no "aperture problem" as there would be in a general 2-D motion analysis system because the displacement analysis is confined to one dimension. We know *a priori* that horizontal displacements in viewing position should lead to pure horizontal displacements in the image. In addition, since the subimages are all formed in a common plane, there is no "keystoning" problem between images as there can be in binocular stereo systems employing nonparallel image planes.

We have used a least-squares gradient technique to measure the displacements [14], [15]. Consider the l-D case: Let $I(x)$ be an image intensity signal; let $I_x(x)$ be the spatial derivative, and let $I_v(x)$ be the derivative with respect to viewing position (dropping the subscript from v for simplicity). Then, the least-squares displacement estimator for a patch of the image is

$$h(x) = \frac{\sum_P I_x(x) I_v(x)}{\sum_P [I_x(x)]^2} \tag{5}$$

where P is the integration patch. A large patch size leads to reduced noise, but it imposes a smoothness on the displacement estimate that blurs over sharp changes. We typically use a patch size of 5×5 to 9×9.

We also assign a confidence estimate at each point as given [14] by the equation

$$c(x) = \sum_P [I_x(x)]^2 . \tag{6}$$

Confidence is high in regions with rapid change (e.g., near lines or edges) but low in regions that are smooth and featureless. Each image pair produces its own displacement map along with an associated confidence map.

The multiple displacement estimates may be combined in a weighted sum, where each estimate is given a weight proportional to its confidence. In a region containing only a horizontal contour, the displacement estimate for the horizontal direction will be of low confidence, whereas the estimate for the vertical direction will be of high confidence; the situation is of course reversed for a vertical contour. The confidence weighted summation rule allows information about horizontal and vertical disparity to be simply and appropriately combined. This combination rule can he formally derived as a leastsquares estimator as is shown in the Appendix.

Alternatively, one could measure orientation in the 4-D space defined by the (x, y) axes of the image plane and the (v_x, v_y) axes of viewing position; the basic concept [8] may be understood by reference to Fig. 5, where the optical arrangement is seen to express depth information as orientation. Just as motion may be analyzed as orientation in (x, y, t) [16], [5], depth may also be analyzed as orientation in (x, y, v_x, v_y). Under reasonable assumptions, this approach also reduces to the estimator described in the Appendix.

We find that it is important to preprocess the image before performing the displacement analysis to remove low-spatial-frequency artifacts that may be caused by nonuniformities in optics, lighting, or sampling. We usually apply a bandpass spatial filter and a local gain control.

Our algorithms sometimes encounter problems with occlusion and disocclusion since the displacement model assumes that the displacement field is smooth, which is not true at an occlusion boundary. More sophisticated algorithms would be required to handle these situations correctly. However, we have not generally found the problems to be grave.

Once the local displacement estimates are derived, they can be converted to depth estimates according to (4).

VII. RESULTS

We constructed a plenoptic camera according to the relay optics scheme shown in Fig. 6(b). The main lens was a 50 mm lens from a 35 mm SLR camera. We made a 2-D lenticular array by crossing a pair of 64 line/in lenticular sheets obtained from Fresnel Optics, Inc. A plano-convex lens served as the field lens. The aerial image formed at the back surface of the lenticular array was reimaged by a Nikon 60 mm macro lens onto a Sony CCD video camera. The resulting image was digitized by a Datacube frame grabber and processed on a Sun 4 workstation. The resolution of the digitized image was 512×480. After subimage analysis, we obtained a set of subimages of approximately 100×100 pixels each.

To reduce aliasing, we placed a weak diffuser in front of the main lens. We tried several devices for the weak diffuser: 1) a so-called diffusion filter manufactured by Hoya and used by portrait photographers to give a soft-focus effect; 2) a pair of crossed lenticular arrays that were optically weakened by the use of index matching fluid with a refractive index very slightly different from that of the lenticular material; 3) a plate of glass lightly sprayed with Suave® Ultrahold hairspray. All three methods worked fairly well, and we used the hairspray diffuser in the experiments to be described here.

A digitized image from the camera is shown in Fig. 7(a). The subject was a pyramid made of Lego® blocks, which provides three distinct heights (like a wedding cake). The tops of the blocks are covered with little bumps, which provide a texture. Each macropixel appears as a small bright patch covering an area of about 5×5 pixels. Fig 7(b) shows an enlargement of a small patch of the image, showing the macropixels more clearly. The mean quantity of light in each macropixel corresponds to the intensity one would record with an ordinary camera with the lens aperture fully open. An individual macropixel is an image of the main lens aperture as seen from a given location in the image plane.

A set of synthesized subimages corresponding to a horizontal sweep of viewing position is shown along the bottom, in Fig. 7(c)-(g); one can generate similar sweeps in other directions. The differences between successive images are difficult to see in this presentation, but when the sequences are shown as animations, the pyramid is seen to rock back and forth.

After applying the displacement analysis in the horizontal and vertical dimensions, we obtain the depth map shown as intensity in Fig. 7(h) and shown in a surface plot in Fig 7(i). The three separate depth planes have been correctly analyzed.

Another example is shown in Fig. 8(a). The subject was a small toy truck, and this image shows one of the extracted subimages. Many regions of this image do not contain enough contour or texture to permit good depth estimates. This limitation is characteristic of all passive optical ranging techniques, such as binocular stereo or motion parallax. In Fig. 8(h), we show the depth estimate obtained in the regions that exceed a threshold confidence level (regions of low confidence are shown in black). In the regions of high confidence, the depth estimates are qualitatively correct. In order to fill in the missing

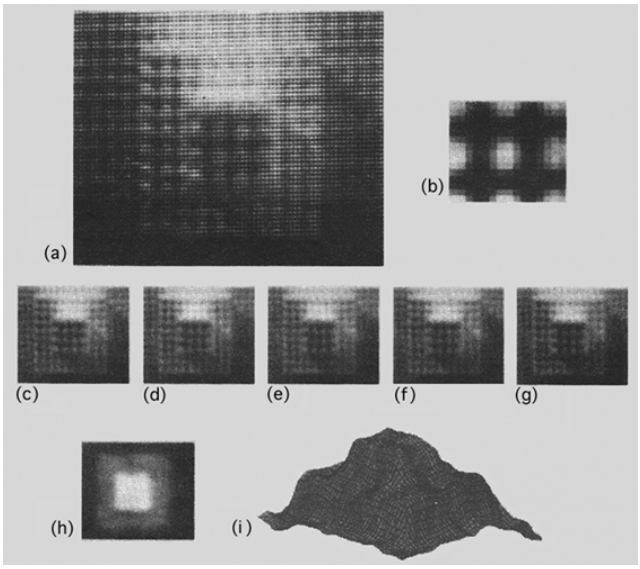


Fig. 7. (a) Digitized image of a lego pyramid, taken with the plenoptic camera. (b) enlarged section of the image, showing the macropixels. (c)-(g) set of subimages taken from virtual viewpoints traversing the lens aperture horizontally; (h) depth map; (i) wire-frame surface plot of the depth map.

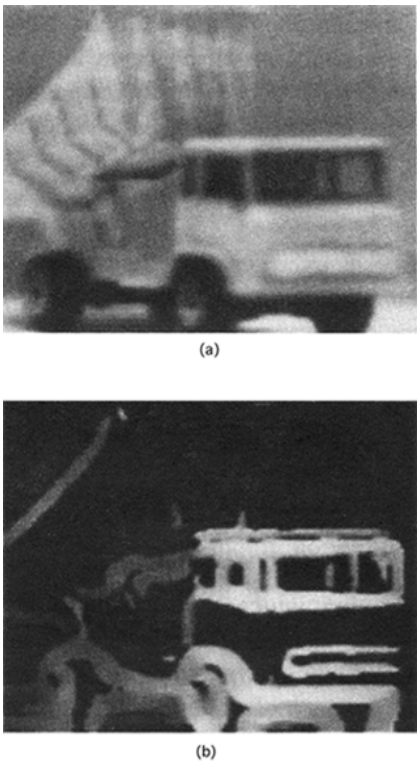


Fig. 8. (a) One subimage of a toy truck; (b) depth map. Values of low confidence are displayed as black.

regions, one would need to use techniques such as those that have been described in other domains of computational vision (e.g., [17]-[l9]).

VIII. CONCLUSIONS

The structure of light that fills the space around an obiect contains a great deal of information that can help characterize the object's 3-D shape. Ordinary camera systems capture only a tiny portion of the information. Binocular stereo systems capture information about two discrete viewpoints; motion parallax systems capture information from a sequence of viewpoints.

The plenoptic camera captures information about the continuum of viewpoints that lie within the aperture of the main camera lens. The structure of the light impinging on the sensor plane is retained by placing a set of miniature cameras there; the cameras can be formed by a pinhole array or a lenticular array. Each tiny camera creates a macropixel that represents the distribution of light within the main lens aperture.

The information so acquired allows one to synthesize images corresponding to different viewpoints within the lens aperture, that is, after the image has been digitized, one can displace the virtual viewing position up, down, left, or right by the application of software. If the optical prefiltering is adequate, there is little or no aliasing in viewing position, and it is possible to move the virtual viewpoint through a continuum of positions. In these circumstances, the correspondence problem is minimized or altogether avoided so that one can use very simple algorithms for disparity analysis. We have used a least-squares gradient procedure to extract information about both the horizontal and vertical parallax, combining the estimates to increase the reliability of the depth estimates.

The system's main limitation is that the stereo baseline is restricted to the size of the lens aperture, which reduces the accuracy with which depth may be resolved.

The system has a number of advantages: it requires only a single camera; it uses both horizontal and vertical disparity; there is no need to establish and maintain calibration between multiple cameras; because the ambiguities of correspondence are minimized, the image processing algorithms can be simple, fast, and robust.

APPENDIX

The least-squares displacement estimate for a 1-D version of the plenoptic camera is described in (5), but the images that are generated by the plenoptic camera are actually described by four parameters: the two spatial dimensions (x, y) and the two viewing dimensions (v_x, v_y). The intensity may be written as $I(x; y, v_x, v_y)$. The least squares displacement estimation for this case would appear at first to involve a difficult 4-D problem, but the constraints imposed by the geometrical optics simplify it greatly.

A displacement in viewing position will lead to an image displacement in a parallel direction. For example, if the viewpoint is moved along the x axis, then features in the image will also move in the x direction; there is no induced parallax in the orthogonal direction (as can occur in binocular systems with nonparallel optical axes). Therefore, we can set up a least-squares estimator as follows.

Let the viewpoint position be (v_x, v_y), and let it be displaced by a small ε distance in direction α. The displacements are then

$$\Delta_x = \varepsilon \cos(\alpha) \tag{7}$$

$$\Delta_y = \varepsilon \sin(\alpha) \tag{8}$$

Displacement of viewpoint leads to a displacement of an image patch by amounts $h\Delta_x$ and $h\Delta_y$. The relationship takes the form

$$I(x,y,v_x,v_y) = I(x - h\Delta_x - h\Delta_y, v_x + \Delta_x, v_y + \Delta_y) \tag{9}$$

or, using shorthand of $c_\alpha = \cos(\alpha)$ and $s_\alpha = \sin(\alpha)$, we can write

$$I(x,y,v_x,v_y) = I(x - h\varepsilon c_\alpha, y - h\varepsilon s_\alpha, v_x + \varepsilon c_\alpha, v_y + \varepsilon s_x). \tag{10}$$

We may then define a squared error to be minimized in order to determine h:

$$E = \int_\alpha \sum_P [I(x,y,v_x,v_y) - I(x - h\varepsilon c_\alpha, y - h\varepsilon s_\alpha, v_x + \varepsilon c_\alpha, v_y + \varepsilon s_\alpha)]^2 \tag{11}$$

where the integral is taken over all displacement directions α from 0 to 2π, and the summation is taken over a 4-D patch P with the axes (x, y, v_x, v_y). Performing a Taylor expansion and retaining the linear terms leads to

$$E \cong \varepsilon \int_\alpha \sum_P [c_\alpha I_x + s_\alpha I_y + h c_\alpha I_{v_x} + h s_\alpha I_{v_y}]^2. \tag{12}$$

Taking the derivative with respect to h and setting it to zero leads to

$$h = \frac{\int_\alpha \sum_P (c_\alpha I_{v_x} + s_\alpha I_{v_v})(c_\alpha I_x + s_\alpha I_y)}{\int_\alpha \sum_P (c_\alpha I_x + s_\alpha I_y)^2} \tag{13}$$

(after cancellation of the common factor, ε). The integral over α may now be performed, and the trigonometric relations of ,c_α, and s_α cause the equation to reduce to

$$h = \frac{\sum_P (I_x I_{v_x} + I_y I_{v_y})}{\sum_P (I_x^2 + I_y^2)} \tag{14}$$

which is the least-squares estimator for the 4-D case. This equation is seen to be a simple extension of (5).

The same estimator can be obtained by performing two l-D displacement estimates and forming a confidence-weighted sum. Let h_x and h_y be the 1-D estimates of displacement in the x and y directions, and let c_x and c_y be the corresponding confidence measures, as described by (5) and (6):

$$h_x = \frac{\sum_P I_x I_{v_x}}{\sum_P I_x^2} \tag{15}$$

$$h_y = \frac{\sum_P I_y I_{v_y}}{\sum_P I_y^2} \tag{16}$$

$$c_x = \sum_P I_x^2 \tag{17}$$

$$c_y = \sum_P I_y^2 \tag{18}$$

Then, the confidence weighted sum of the two l-D estimates is

$$h_{cw} = \frac{h_x c_x + h_y c_y}{c_x + c_y} \tag{19}$$

$$= \frac{\sum_P (I_x I_{v_x} + I_y I_{v_y})}{\sum_P (I_x^2 + I_y^2)}. \tag{20}$$

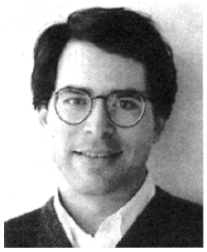
This is identical to the least-squares estimator of (14).

ACKNOWLEDGMENT

We thank S. Benton, A. Pentland, and E. Simoncelli for useful discussions.

REFERENCES

[1] J. P. Richter, ed., *The Notebooks of Leonardo da Vinci*. New York: Dover, 1970, p. 39, vol. 1.
[2] W. L. L. Grimson, *From Images to Surfaces: A Computational Study of the Human Early Visual System*. Cambridge, MA: MIT, 1981.
[3] M. Ito and A. Ishii, "Three view stereo analysis," *IEEE Trans. Patt. Anal. Machine Intell*., vol. PAMI-8, pp. 524-531, 1986.
[4] N. Ayache and F. Lustman, "Fast and reliable passive trinocular stereo vision," in *Proc. ICCV*, pp. 422-427.
[5] R. C. Bolles, H. H. Baker, and D. H. Marimont. "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comp. Vis*., vol. 1, pp. 7-55, 1987.
[6] A. P. Pentland, "A new sense for depth of field," *IEEE Trans. Patt. Anal. Machine Intell*., vol. PAMI-9, pp. 523-531, 19-37.
[7] V. M. Bove, Jr., "Probabilistic method for integrating multiple sources of range data," *J. Opt. Soc Amer. A*, vol. 7, pp. 2193-2207, 1990.
[8] E. H. Adelson and J R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing* (M. Landy and J. A. Movshon, Eds.). Cambridge, MA: MIT Press, 1991.
[9] J. J. Gibson, *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin, 1966.
[10] G. Lippmann, "Epreuves reversibles donnant la sensation du relief." *J. Phys.7*, pp.821-825, 1908.
[11] H. E. Ives, "Parallax panoramagrams made with a large diameter lens," *J. Opt. Soc. Amer.*, vol. 20, pp 332-342, 1930.
[12] T. Okoshi. *Three Dimensional Imaging Techniques*. New York Academic, 1976.
[13] S. F. Ray, *Applied Photographic Optics: Imaging Systems for Photography, Film, and Video.* Boston: Focal, 1988.
[14] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Image Understanding Workshop*, pp. 121-130.
[15] C. Caffario and F. Rocca, "Methods for measuring small displacements of television images," *IEEE- Trans. Inform. Theory*, vol. IT-22, pp. 573-579, 1976.
[16] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer*., vol. A2, pp. 284-299, l985.
[17] W. E. L. Grimson, "An implementation of a computational theory of visual surface interpolation," *Comput. Vision Graphics, Image Processing*, vol. 2, pp. 39-69, 1983.
[18] D. Terzopoulos, "The computation of visible surface representations," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 10, pp. 417-439, 1988.
[l9] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory ," *Nature*, vol. 317, pp. 314-319, 1985.



Edward H. Adelson received the B.A. degree in physics and Philosophy from Yale University and the Ph.D. degree in experimental psychology from the University of Michigan.

He is currently an Associate Professor of Vision Science at the Media Laboratory and the Department of Brain and Cognitive Science at the Massachusetts Institute of Technology. He has published numerous papers in the fields of human visual perception, visual neurophysiology, computational vision, image processing, image communications, and computer graphics; he also holds numerous patents.

Dr. Adelson is a Fellow of the Optical Society of America and received that society's Adolph Lomb Medal in 1984.

John Y. A. Wang received the S.B. and S.M. degree in electrical engineering; from the Massachusetts Institute of Technology in 1987.

His research experience includes work in cir-cuit design at MIT, Texas Instruments, and Hughes Aircraft. Currently, he is a research assistant at the MIT Media Laboratory Vision and Modeling Group, working on the Ph.D. degree in electrical engineering. His recent research is in human and machine vision with special interest in motion analysis and 3-D sensors.

Mr. Wang is a member of Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.