A Perspective on Range Finding Techniques for Computer Vision

R. A. JARVIS

Abstract—**In recent times a great deal of interest has been shown, amongst the computer vision and robotics research community, in the acquisition of range data for supporting scene analysis leading to remote (noncontact) determination of configurations and space filling extents of three-dimensional object assemblages. This paper surveys a variety of approaches to generalized range finding and presents a perspective on their applicability and shortcomings in the context of computer vision studies.**

Index Terms—**Computer vision, range finding.**

I. Introduction

IT IS well documented in the psychological literature [1]-[3] that humans use a great variety of vision-based depth cues, combinations from this repertoire often serving as confirming strategies with various weighting factors depending upon the visual circumstances. These cues include texture gradient, size perspective (diminution of size with distance), binocular perspective (inward turning eye muscle feedback and stereo disparity), motion parallax, aerial perspective (haziness, etc. associated with distance), relative upward locations in the visual field, occlusion effects, outline continuity (complete objects look closer), and surface shading variations. In difficult circumstances, various of these cues provide evidence of feasible interpretations; resolving ambiguity when it exists depends not only on the sensory information immediately available but also on previously formed precepts and consequent expectation factors. It is interesting to note that binocular convergence (related to muscle driven inward turning of the eyes) adjusts the scale of the stereo disparity system. Of the human vision depth cues, from a geometric point of view, convergence and disparity are unambiguously related to distance whereas, without movement, perspective depth cues are intrinsically ambiguous as can be easily demonstrated by the Ames room illusions [2]. Thus, convergence and disparity are good candidates for depth estimation in computer vision studies. However, the apparatus available to support depth estimation in machine vision can extend beyond anthropomorphically based cues; these will be detailed later.

Categorization of the various types of range finding techniques is useful in providing a structure for detailed discussion. Direct and active range finding includes ultrasonic and light time-of-flight estimation and triangulation systems. All involve a controlled energy beam and reflected energy detection. Most other range finding methods can be generally classified as image based, but further refinement is helpful. Passive (or relatively so) monocular image-based range finding includes texture gradient analysis, photometric methods (surface normals from reflectance), occlusion effects, size constancy, and focusing methods. Contrived lighting approaches include striped and grid lighting, patterned lighting, and Moiré fringe analysis. This leaves methods based either on motion or multiple relative positions of camera or scene; these include reconstruction from multiple views, stereo disparity, retinal flow, and other motion related techniques. Most of these are really geometric triangulation systems of one kind or another. In fact, almost every circumstance that includes at least two views of a scene is potentially exploitable as a range finding mechanism of a triangulation kind; equally true is the drawback that all triangulation methods potentially suffer from the problem of "missing parts" not included in more than one view. In contrast, a coaxial source/detector arrangement for a time-of-flight laser range finder is not subject to this malady.

In general, passive methods have a wider range of applicability since no artificial source of energy is involved and natural outdoor scenes (lit by the sun) fall within this category. However, for complex scenes, the degree of ambiguity in need of resolution is likely to be higher if intrusive methods such as using ultrasonics, laser-beams, and striped lighting are not applicable. On the other hand, ranging methods using structured light sources or time of flight measuring devices, although perhaps contributing little to our understanding of human vision, are certainly acceptable in indoor factory environments where these active approaches are consistent with other instrumentation methodologies.

It will be assumed that the type of range finding required is that which results in a "rangepic," an array of distance estimates from a known point or plane with adjacency constraints corresponding to those of two-dimensional intensity imagery. Not only does this allow a direct correspondence to be made with intensity imagery, but also indicates the amount of information associated with the results and puts some time constraints on range data acquisition if robotic manipulation is to be carried out in a reasonable time span. That a "rangepic" is in fact an "image" is more easily argued in this context; however, a purist might argue that the use of directly acquired range data puts this analysis outside the scope of legitimate computer vision. That both intensity and range data can be remotely acquired to plan manipulation trajectories, is, however, most valuable. The combination of range and intensity

Manuscript received April 10, 1981; revised September 2, 1982.
The author is with the Department of Computer Science, Australian National University, Canberra, Australia.

0162-8828/83/0300-0122$01.00

data to this end is worthy of careful study because of the high potential of resolving scene interpretation ambiguities in this way, without heavy dependence on semantically derived guidance which might severely restrict the breadth of applicability.

The paper first deals with two contrived lighting ranging methods. Then follows coverage of the monocular passive techniques of relative range from occlusion cues, range from texture gradient, range from focusing, and surface orientation from brightness. The multiple camera position techniques of stereo disparity and camera motion into the scene are then addressed. This is followed by a section on Moiré fringe range contouring, which, although in the category of contrived lighting methods, is presented later in the paper both because more specialized instrumentation is involved and because a photographic intermediate step makes it unsuitable for real-time range analysis. Then follows the artificial beam energy source methods which range to one point at a time: simple triangulation active ranging, ultrasonic and laser time-of-flight active methods, and a streak camera approach which provides an interesting and fast method for measuring light transit times with great accuracy.

Although some of the methods presented may seem to have considerable drawbacks which could throw doubt on why they are included, it was felt that a wide representative spread of ranging approaches should be described to provoke thought and development in this important, relatively new field, bearing in mind that new technologies could change the feasibility status of various methods for particular applications.

II. Contrived Lighting Range Finding

In many laboratory situations where experiments in computer vision are intended to have applications in the component handling, inspection and assembly industry, special lighting effects to both reduce the computational complexity, and improve the reliability of 3-D object analysis is entirely acceptable. That similar methods are not applicable in generalized scene analysis, particularly out of doors, is of no great concern. This class of range finding method involves illuminating the scene with controlled lighting and interpreting the pattern of the projection in terms of the surface geometry of the objects.

A. Striped Lighting (See Fig. 1)

Here the scene is lit by a sheet of light usually produced with a laser beam light source and a cylindrical lens, but projecting a slit using a standard slide projector is also feasible. This sheet of light is scanned across the scene, producing a single light stripe for each position. When the light source is displaced from a viewing TV camera, the camera view of the stripe shows displacements along a stripe which are proportional to depth; a kink indicates a change of plane and a discontinuity a physical gap between surfaces. The proportionality constant between the beam displacement and depth is dependent upon the displacement of the source from the camera so that more accurate depth measurements can be made with larger displacements; however, larger parts of the scene viewable from the source position (lightable) are not seen from the TV camera—the depth of such portions cannot be measured in this way. High source energies permit operation in normal ambient lighting conditions.

It is simpler to analyze one stripe at a time in the image since line identification for tracing purposes becomes difficult with multiple lines, particularly when discontinuities occur. Using specially identifiable adjacent stripes using dashes or color coding could reduce the number of images requiring analysis. In [4] Shirai and Suwa use a simple stripe lighting scheme with a rotating slit projector and TV camera to recognize polyhedral shapes. The fact that the projected lines are equispaced, parallel, and straight on planar surfaces are taken advantage of in reducing the computational complexity; only the endpoints of each straight line segment are found and line grouping procedures are used to identify distinct planar surfaces. The geometry of the relationship between depth and displacement of a point on a stripe is not confounded by non-planar surfaced objects but point by point depth analysis involving one binary image for each stripe is expensive computationally. Furthermore, if a rangepic on a uniform grid is required, some extrapolation and interpolation calculations will be required.

In [5] Agin and Binford describe a laser ranging system capable of moving a sheet of light of controllable orientation across the scene. A helium neon laser emitting 35 mW of red light at a wavelength of 6328 Å was thought adequate for use with the vidicon camera used. Their aim was to derive descriptions of curved objects based on the generalized cylinder model [6]. The stripe line from an image at each of a number of positions of a rotating mirror were analyzed, the orientation of the sheet of light rotated by 90°, and a second mirror scan sweep taken. This process resulted in data for an overlapping grid of laser lines covering the scene. Some 5-10 min were involved in the process. Range data are derivable from these data since the relative position of the laser beam with respect to the camera can be determined by a calibration process. The disadvantages of the apparatus as cited by the authors include slowness of data collection and low-level processing, the monochromaticity of the laser source in restricting the hue of the objects to the scanned, and the hazards present during the use of lasers in an uncontrolled environment.

The work reported by Popplestone *et al.* [7] differs from Shirai's in that it deals with both cylindrical and plane surfaces and from both Shirai's and Agin's work in the development of body models specifically suited to solving juxtaposition problems in automatic assembly. The stripe analysis hardware takes advantage of the facts that 1) a nearly vertical stripe intersects each horizontal scan of a TV camera only once and that 2) since each TV line scan is at constant speed, the time from the start of that line at which the video signal indicates an intersection "blip" is proportional to the distance of the stripe from the left edge of the image. The stripe finder electronics returns the relevant timing data to the controlling minicomputer which is fast enough to collect the data for one complete stripe in one TV frame time (1/50 s). Hardware details are given in [8].

It is evident that without special video signal timing electronics, many frames, each with little relevant information,

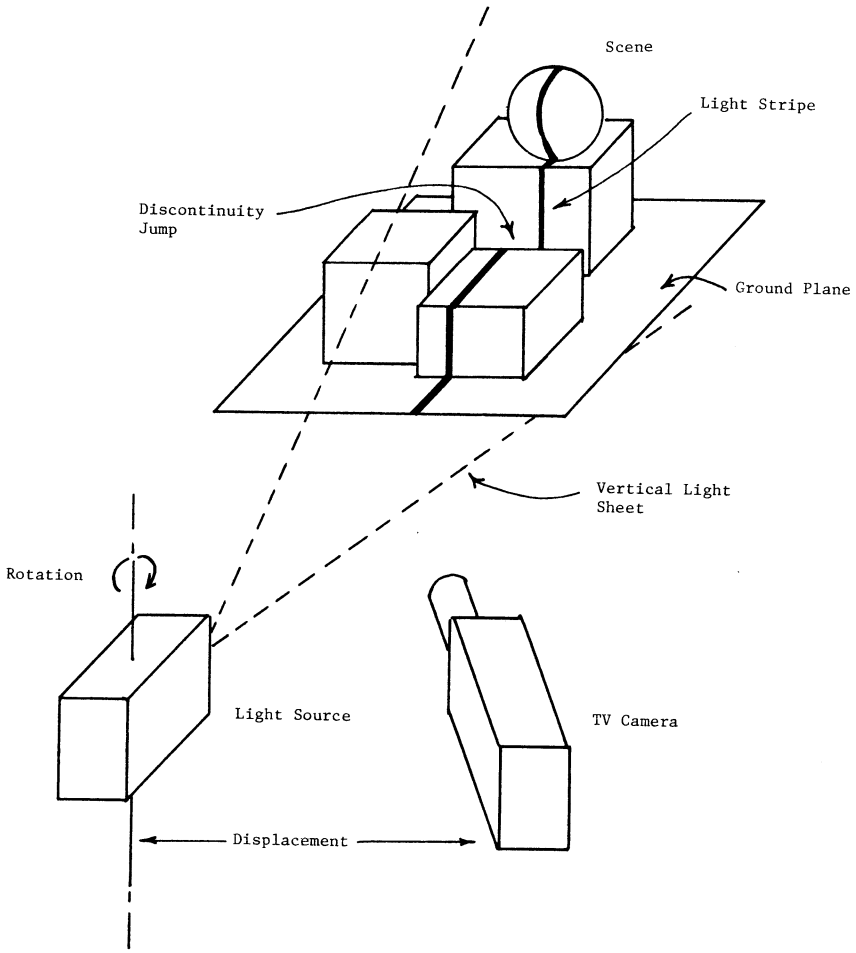


Fig. 1. Striped lighting apparatus.

have to be stored and analyzed, leading to a time-consuming range extraction process. Röcker and Kiessling [9] discuss this problem along with difficulties associated with other ranging techniques. In particular, they point out that, if more is to be extracted from a single image frame by using parallel grid illuminations, the strike identification problem causes a number of restrictions related to the following.

1) The image should contain parts of the supporting plane surface.

2) Shadows cause line interruptions.

3) Top surface lines should be distinguishable from ground plane lines.

4) Scenes with more than one object should not have hidden object planes.

This last point would appear to affect nearly all types of ranging and all image based scene analysis, for that matter.

B. Grid Coding

Will and Pennington [10] describe a method by which the locations and orientations of planar areas of polyhedral solids are extracted through linear frequency domain filtering applied to images of scenes illuminated by a high contrast rectangular grid of lines. Edges are defined through the intersections of the extracted planes. Fast Fourier 2-D transforms are used for rapid computation and segmentation in the Fourier domain used to identify the planes in the grid coded image. Once again the TV camera is offset from the illumination source. The transformation matrix which would restore the individual distorted squares back to their original form in the projected grid contains the local surface normal directions; but here Fourier domain analysis is used instead. The grid coded planar areas map into a 2-D Fourier transform which is a crossed set of harmonically related delta functions in the spatial frequency domain. Separation of the delta functions identifying the planes is equivalent to bandpass filtering; the inverse transform is a reconstruction of the isolated planes in the image domain. Higher level processing can then deduce the object structures in terms of the identified planes. A filter consisting of a 1° sector of a circle with radial direction all-pass response was applied to the 2-D Fourier spectrum to produce an energy versus angle function, the peaks of which are associated with individual planes; a set of filters was designed to straddle each peak; each filter then passed those parts of the 2-D Fourier spectrum corresponding to the individual planes; inverse transforms reconstruct the planes in the image domain.

Towards the end of the paper, an interesting alternative approach is mentioned. If a photographic camera is moved transversely to the dominant direction to the scene, each image point gives rise to a streak of length inversely propor-

tional to range. These streaks are coded by shuttering the camera at constant intervals; each is encoded as an array of points whose period is proportional to range. This periodicity can be detected in the Fourier domain. This approach would seem to be time-consuming and inconvenient, particularly if photographic processing is involved. Furthermore, only distinct points in the scene would give rise to clear modulated streaks on the image plane; range to other parts would need to be calculated on the basis of assumptions of the shapes of surfaces and other clues.

The Fourier analysis approach to 3-D computer vision would be feasible for robot guidance only if the computations involved could be completed quickly enough, perhaps with array processing support.

III. Relative Range from Occlusion Cues

Rosenburg *et al.* [11] have developed a technique for computing the relative relationships of "in-front-of," "behind," and "equidistant" using heuristic evidence of occlusion in monocular color imagery. A relaxation labeling [12], [13] process is used to produce a depth map which is used to test the consistency of a depth graph derived from occlusion cues. The scheme functions without domain-specific restrictions. A segmented image is used as input—each region is assumed to be distinguished from adjacent regions on the basis of the primary features of color and texture. It is assumed that if some of these regions represent only parts of objects, this is purely the result of occlusion effects. Occlusion evidence is obtained by examining clusters of adjacent regions, each cluster being evaluated in terms of six different occlusion cases (see Fig. 2) ordered in decreasing evidence of occlusion. In Fig. 2(a) region A is totally contained within region B. This represents the strongest evidence of occlusion, there being no breaks in the occluded region. In Fig. 2(b) the occluded object is broken somewhat. Region A is surrounded by region B on at least 75 percent of its perimeter and the fragment region C has the same primary feature properties as region B. It is likely that B and C are parts of the same object. Removing region C gives rise to Fig. 2(c) which is ranked below (b). In Fig. 2(d), the reduction of the occluded region B weakens the clue further—A is surrounded by B on at least 50 percent of its perimeter. Again the additional area C with feature properties similar to B strengthens the occlusion hypothesis a little. In Fig. 2(e), the lack of region C weakens the hypothesis. In Fig. 2(f) the extent of the adjacency of region B is reduced, A is surrounded by B on at least 25 percent of its perimeter; once again C can help.

Where there are distinct occlusion related groupings with no occlusion clues between the groupings, relative depth relationships between the groups cannot be determined in this way. Probabilistic relaxation labeling [12] is used both to resolve/reduce possible contradictions in local occlusion data based on the six classes given above and to establish the "equidistant" relationship. Labels are attached to each region indicating hypotheses about depth level of the underlying physical objects. If N depths are used (depth level 1 is foreground and depth level N is background) and λ_α, $\alpha = 1, 2, \cdots, N$ are the labels used, $p_i(\lambda_\alpha)$ is the probability that λ_α is the correct depth level for region i. The total number of regions in occlusion clusters is taken as an upper bound on N. Initially all $p_i(\lambda)$ are set to $1/N$.

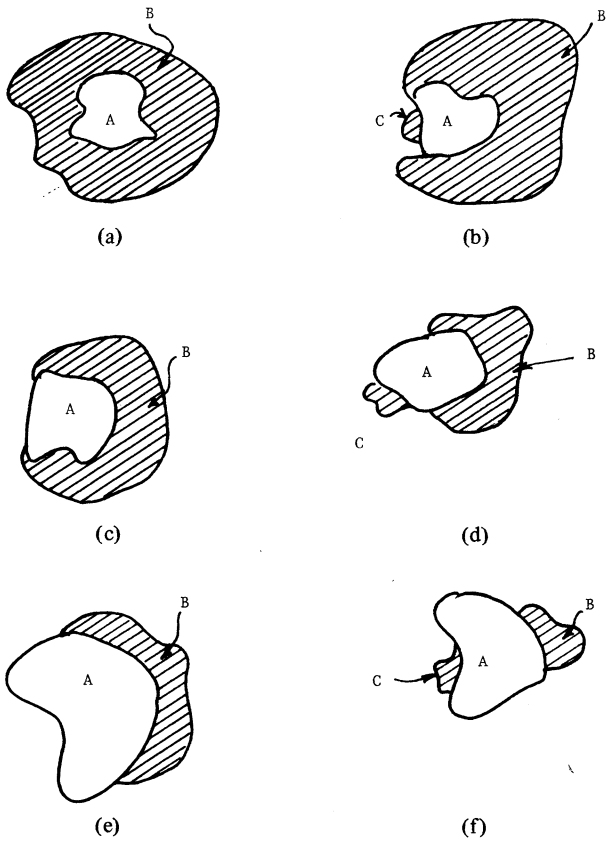


Fig. 2. Occlusion cases.

These probabilities are updated using

$$p_i^{k+1}(\lambda) = \frac{p_i^k(\lambda)\,[1 + q_i^k(\lambda)]^\gamma}{\sum_\lambda p_i^k(\lambda)\,[1 + q_i^k(\lambda)]^\gamma}$$

maintaining

$$\sum_{\alpha=1}^{N} p_i^{k+1}(\lambda_\alpha) = 1 \qquad \text{for each } i$$

where we have the following.

1) $$q_i^k(\lambda) = \sum_{j \in \text{NEIGH}(i)} c_{ij} \sum_{\alpha=1}^{N} r_{ij}(\lambda, \lambda'_\alpha)\, p_j^k(\lambda'_\alpha).$$

2) k is the iteration number and $\gamma > 1$ an accelerating factor. NEIGH(i) is the adjacency neighborhood set of i.

3) $r_{ij}(\lambda, \lambda')$ is the compatibility (or consistency) between label λ on region i and label λ' on region j. In this example,

$$r_{ij}(\lambda, \lambda') = -1 \quad \text{if } \lambda \geq \lambda'$$
$$= +1 \quad \text{if } \lambda < \lambda'$$

if i is an occluding region and j the corresponding occluded region.

4) c_{ij} is relative certainty of inferences attached to each of the six occlusion cases for region pair i, j. The assignment used

in this paper is

$$c_{ij} = \begin{cases} 0.6 & \text{for Fig. 2(a)} \\ 0.5 & \text{for Fig. 2(b)} \\ 0.4 & \text{for Fig. 2(c)} \\ 0.3 & \text{for Fig. 2(d)} \\ 0.2 & \text{for Fig. 2(e)} \\ 0.1 & \text{for Fig. 2(f).} \end{cases}$$

Convergence usually occurs after applying the update function a medium number of times (in paper, 22 iterations used for example presented). γ can be changed at any iteration if it seems helpful.

The final step is the determination of depth relationships of each region with each other region; this process is not presented here.

The most glaring weakness in this approach is the restraint of correct segmentation of the scene in the first place. In many practical situations one would wish that range information would help resolve segmentation ambiguities, not range information itself to depend upon the lack of these. Total reliance on occlusion cues is also a weakness in that object groups not linked by occlusion relationships cannot be relatively placed in the depth map. However, the approach is most ingeneous and deserving of attention, particularly because of its monocular application.

IV. Depth from Texture Gradient

Texture gradient refers to the increasing fineness of visual texture with depth observed when viewing a 2-D image of a 3-D scene containing approximately uniformly textured planes or objects. Gibson [14] placed considerable emphasis on this effect in terms of human depth cues, particularly when associated with the ground plane. It is intrinsically a monocluar phenomenon particularly useful in range analysis on natural outdoor scenes where uniform visual texture is a dominant manifestation.

Bajcsy and Lieberman [15] have developed a method of measuring texture gradients in the domain of natural outdoor scenes based on Fourier descriptors which are claimed to vary in a manner consistent with surface geometries in three dimensions. The Gibson [14] point of view is supported—surfaces are the primary objects of the visual world; these reflect light, some of which is projected on the retina. The basic surface classes are longitudinal (parallel to line of sight) and frontal (transverse to line of sight); longitudinal surfaces are associated with distance perception.

The Bajcsy and Lieberman texture operator is developed as follows.

The 2-D discrete Fourier transform of a digitized image window considered as a real function $g(x, y)$ of two spatial integer variables x, y is

$$F(n, m) = \frac{1}{p^2} \sum_{x=0}^{p-1} \sum_{y=0}^{p-1} g(x, y) \exp\left[-2\pi i(xn + ym)/p\right]$$

where p is the dimension of the square image window array ($0 \leqslant x, y \leqslant p$, all integers).

The power spectrum is

$$P(n, m) = [F_{R_e}^2(n, m) + F_{I_m}^2(n, m)]^{1/2}$$

where $F_{R_e}(n, m)$ and $F_{I_m}(n, m)$ are the real and complex parts, respectively, of $F(n, m)$. The phase spectrum is

$$\psi(n, m) = \arctan\left[F_{I_m}(n, m)/F_{R_e}(n, m)\right].$$

The power spectrum, being invarient to translation but not rotation preserves the visual pattern directionality of the image; the phase spectrum contains position information in the image and is not relevant for texture cues. Transforming the power spectrum from Cartesian (n, m) to polar (r, ϕ) coordinates aids extraction of directionality information. In each direction ϕ, $P(r, \phi)$ is a one-dimensional function $P_\phi(r)$; for each frequency, r, $P(r, \phi)$ is a one-dimensional function, $P_r(\phi)$. The paper is not clear in distinguishing $P_\phi(r)$ functions for each ϕ from a single function $P(r)$ formed by integrating over ϕ nor in distinguishing between $P_r(\phi)$ for each r and a single function $P(\phi)$ formed by integrating over r.

In [15] in fact, an example is given where $P(\phi)$ is calculated by summing the energy spectrum $P(n, m)$ in equiangular sectors and then finding $P(r)$ for each direction indicated by peaks in $P(\phi)$ by summing the energy spectrum $P(n, m)$ within the associated ϕ sector and within rectangular annuluses of radius r. The nomenclature used is not consistent.

Peaks in $P(\phi)$ indicate texture directionality—a few distinct peaks indicate strong directionality properties, a uniform function indicates nondirectionality of texture. A significant peak is taken as one higher than the mean by 1.5 times the standard deviation. In the nondirectional texture case a uniform $P(r)$ indicates a noisy texture and a peaky $P(r)$ a blob-like texture. A lack of texture (smooth image) gives rise to a large $P(r)$ at $r = 0$ (zero frequency component).

The quantitative components used for the texture descriptor include:

1) average gray value (zero frequency constant),

2) the number and angle values of prime textural directionalities,

3) maximum power corresponding to each prime directionality,

4) the r corresponding to the maximum power in each prime direction, and

5) their corresponding spatial frequencies and wavelength.

Qualitative components include:

1) texture class (bloblike, monodirectional, noisy, homogeneous, etc.),

2) contrast (sharp, medium, weak),

3) brightness (bright, dark),

4) granulation (large, medium, small).

The depths derivable from texture gradient are only relative unless the actual size of the texture element is known as a basis of calibration.

Using the geometric model of Fig. 3, l_i is the texture element size as projected on the image plane, Y_i the center of the window in which l_i is found. The ground plane texture elements are all the same size (say = t). Y_A, Y_B, Y_C are Y value image projections of points A, B, C which are on the ground plane. The relative distances of A, B, and C to the image plane are to be determined.

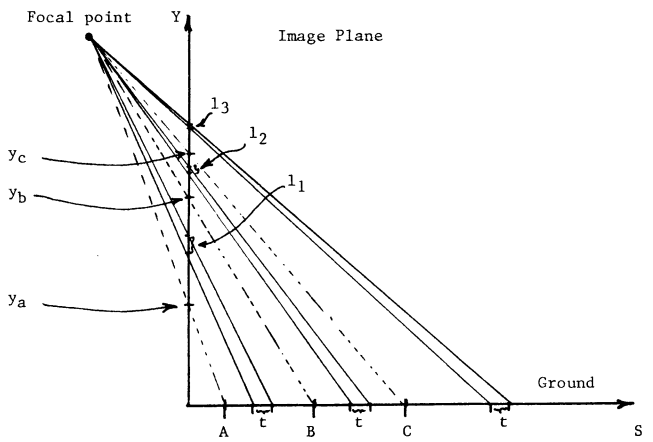


Fig. 3. Texture gradient geometric model.

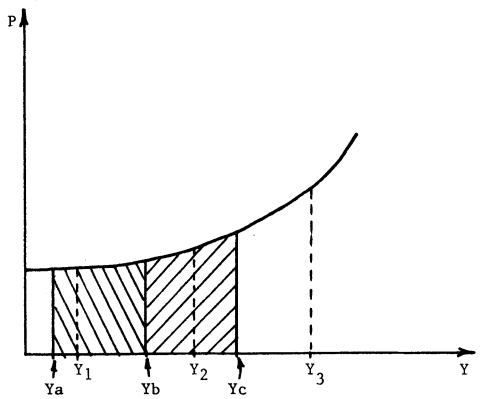


Fig. 4. Projection function.

$P = (1/k_1)\,(ds/dy)$ is a projection function which indicates how distance on the ground plane is related to distance on the image plane where k_1 is a proportionality constant depending on geometric system parameters. Distance in 3-D space is given by

$$S = k_1 \int P\, dy.$$

The texture descriptors extracted earlier can be used to produce a form of P,

$$P_i(Y_i) = k_2 t / l_i(Y_i)$$

where Y_i identifies the associated windows in the image plane and k_2 is another constant

$$\frac{\text{distance } AC}{\text{distance } AB} = \frac{\int_{YA}^{YC} P^*\, dy}{\int_{YA}^{YB} P^*\, dy}$$

where P^* is a curve fit approximation to P.

This is shown in Fig. 4. The l_i's are the texture wavelengths or elements sizes measured in adjacent windows along the vertical direction of the image. Thus relative distances can be found in terms of texture gradient without needing to know focal length, height above ground, etc.

Window size requires careful consideration since relativity to texture coarseness is needed to give reliable texture features.

This method of relative range measurement would seem to have several drawbacks. Firstly, the regions of the image over which the texture features are to be extracted must be uniformly textured in the 3-D sense. Prior segmentation is required. Secondly, application is restricted to highly textured scenes. Thirdly, computational cost, despite use of fast FFT algorithms, would be high.

Other texture coarseness measures [16], [17] might be substituted in the above method to derive relative depths.

V. Range from Focusing

Knowledge of the focal length and focal plane to image plane distances permits evaluation of focal plane to object distance (range) for components of the 3-D scene in sharp focus. The sharpness of focus needs to be measured on windows on the image over a range of lens positions to determine the range of the corresponding components of the scene. Prior segmentation is not required, but sufficient visual "business" is required to enable sharp focus. Large lens apertures shorten the depth of focus and enhance focus position discrimination of objects at different ranges. Horn [18] provides some technical details on focusing relationships and Jarvis [19] suggests some simple computational formulas for sharpness of focus evaluation. The method is essentially simple and a direct calibration procedure can be used to associate lens positions for various ranges in focus, thus obviating the need for mathematical derivation of this function. The method becomes increasingly inaccurate with range (see Fig. 5).

Jarvis [19] suggests the following focus sharpness measures, chosen on the basis of computational simplicity, effectiveness, consistency and possible direct hardware implementation:

1) entropy (comentropy) $= E = -\sum_x P(x) \ln P(x)$.

2) variance $= V = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$; $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

3) sum modulus difference $=$ SMD $= \sum_{i=2}^{N} |x_i - x_{i-1}|$.

For each window on the image one need only find the lens position (and thus range) which maximizes these functions; no texture details are required and no absolute values are important.

Instead of considering each of a set of rectangular grid based windows in the image over a set of lens positions, it is also possible, for any one complete image, to identify those portions which are in focus and thus derive the range of the corresponding objects.

Once again, as with the texture gradient approach, the range to visually homogeneous regions of the image cannot be determined directly. However, only one camera position is involved and no special apparatus (except perhaps a computer controlled motor to adjust the lens position) is required in addition to standard digitization equipment if the focus sharpness calculations are to be computed rather than determined with

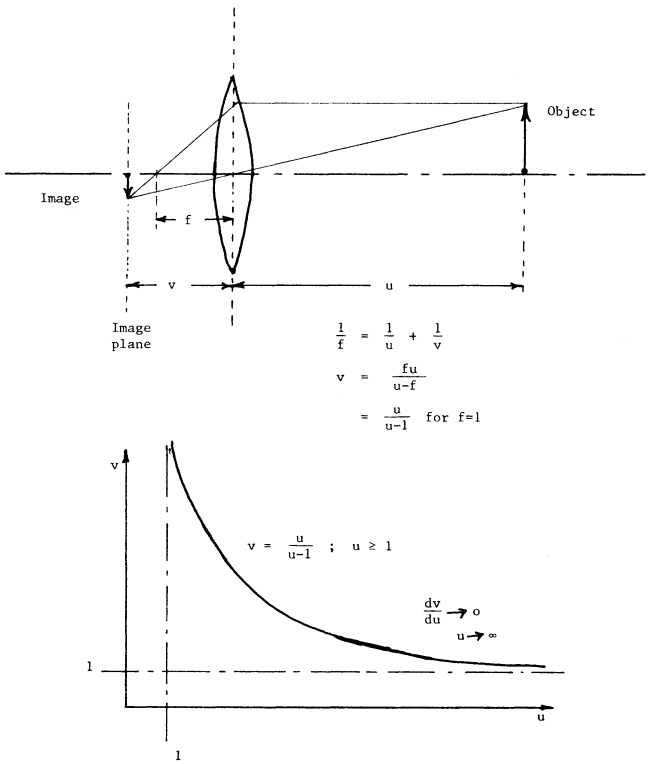


Fig. 5. Depth from focus principle.

specialized analog electronic hardware. In the former case, calculations, although straightforward, could be lengthy.

VI. Surface Orientation from Image Brightness

As early as 1970 Horn [20] raised issues concerning the recovery of surface shape from shading information in the image. In more recent times there has been a renewal of interest in this approach to scene analysis as it represents a generalized analysis strategy independent of domain specific restrictions. Only the recent work by Ikeuchi and Horn [21] will be briefly presented here as representative of this approach to surface orientation recovery. As mentioned earlier, surface orientation permits relative range information over parts of the scene to be calculated by integration; discontinuities frustrate absolute range determinations over the entire scene. Central to the method is the concept of a reflectance map which captures the relationship between image intensity (shading) and surface orientation.

Denoting the slope components of a surface patch as p, q

$$p = \partial z/\partial x$$

$$q = \partial z/\partial y$$

where z is the depth coordinate and the brightness distribution on the surface gradient space, $R(p, q)$, is called the "reflectance map."

The dominant image brightness relationship to reflectance map is

$$E_p(x, y) = R(p, q, p_s, q_s)$$

where

$E_p(x, y)$	is the image irradiance in the image plane at (x, y) (orthographic projection assumed)
(p_s, q_s)	are the direction components of the light source
$R(p, q, p_s, q_s)$	is the reflectance map function defined over surface orientation and light source position.

The central mechanism of surface normal recovery is to calculate the $R(p, q, p_s, q_s)$ map off-line for the surface material of the scene and to determine p, q for each x, y image point from solving a set of $E_p(x, y) = R(p, q, p_s, q_s)$ equations with different light source positions but with camera and scene stationary. The paper deals with surfaces with high specular reflectance properties which the authors suggest are typical of industrial objects, but the general approach is not restricted to these types of surfaces. A considerable amount of off-line computation of the reflectance map function is required but on-line scene analysis is largely by table look-up which is rapid. Surfaces with indirect illuminations from adjacent components cannot be analyzed reliably in this way and the method would be restricted to objects of the one type of surface for which the reflectance map has been calculated off-line. The accuracy with which image intensity can be evaluated would also seem critical to the method. In an industrial hand/eye coordination system these restrictions may not be prohibitive.

VII. Range From Stereo Disparity

Stereo disparity refers to the phenomenon by which the image of a 3-D object point shifts as the camera is moved laterally to the depth coordinate axis. For two such camera positions, simple geometry indicates that the image displacement (disparity) is inversely proportional to depth as measured from the camera (see Fig. 6). The image of a point at an infinite distance along the optical axis can be used as a reference position in both images. Disparity relative to a line through this reference point on the image at right angles to the camera shift direction is inversely proportional to depth. (In the limit, the image of the infinitely distant point does not shift at all.)

It is necessary to establish correspondence or matching of points between the two images to derive the depth relationship. If this correspondence is to be determined from the image data there must be sufficient visual information at the matching points to establish a unique pairing relationship. Two basic problems arise in relation to this requirement. The first arises at parts of the image where uniformity of intensity or color makes matching impossible, the second when the image of some part of the scene appears in only one view of a stereo pair because of occlusion effects (the missing parts problem) or because of the limited field of view captured on the images. The further apart the two camera positions, the potentially more accurate the disparity depth calculation—but the more prevalent the missing parts problem and smaller the field of view overlap.

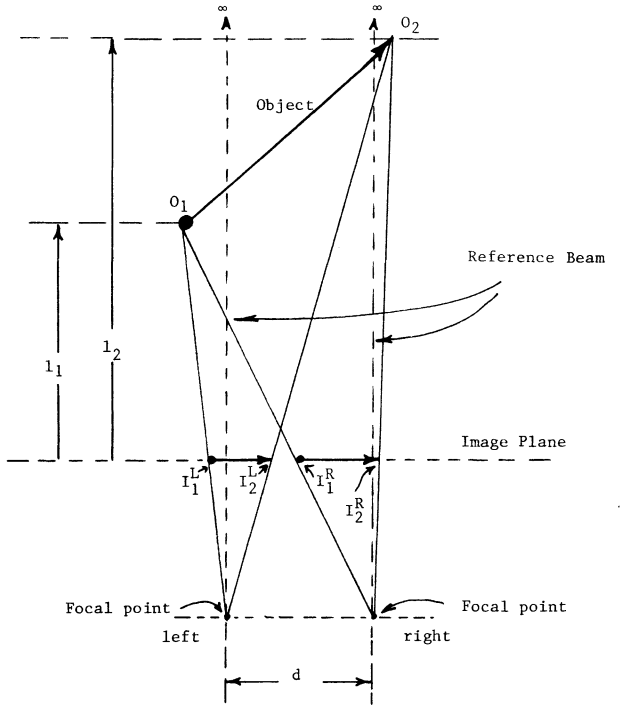


Fig. 6. Stereo disparity geometry.

If the correspondence problem is to be tackled using correlation maximization over windows of the image pair, the correlation shift need only be in the direction of the camera movement axis if this is known.

There is quite a lot of literature on the stereo disparity range finding problem and [22], [24]-[26] are only a sample from this field.

The solution of the correspondence may be effectively sought over the entire overlap areas of the stereo pair of images if the scene is visually "busy" over most of its imaged extent. When large areas of the image are relatively featureless, correlation window matching attempts would prove futile and it is more reliable and expedient to preselect portions for matching on the basis of scene "busyness" measure of a textural or line structure sensitive type. Hopefully, these preselected areas are strategically sufficiently well placed to allow extrapolation and interpolation based depth estimates to be reliably made for unrepresented portions of the scene. Certainly in man-made environments with planar faced solids, this approach is likely to be fruitful. The "busyness" measures that are likely to be suitable include many of the same measures that relate to texture quantification [15]-[17] and also those suggested earlier which have proven useful [19] as focus sharpness measures.

Levine *et al.* [22] apply the following correlation measure over $(2u+1) \times (2v+1)$ windows of the stereo image pair $A(i, j)$, $B(i, j)$; the windows are centered off some point (i, j):

$$\phi(p) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \cdot \frac{\{[A(\xi,\eta)\, B(\xi,\eta+p)] - \mu_A(i,j)\, \mu_B(i,j+p)\}}{\sigma_A(i,j)\, \sigma_B(i,j+p)}$$

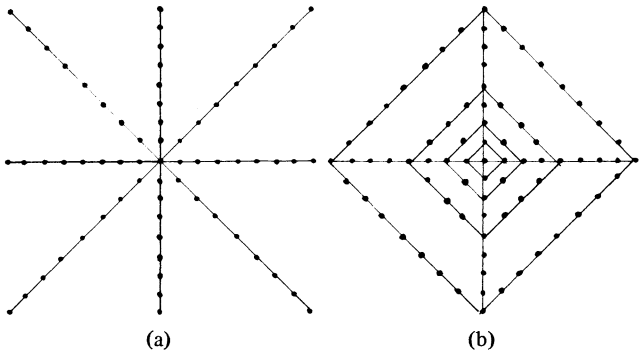


Fig. 7. Correlation masks.

and

$$\mu_B(i, j+p) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j+p-v}^{j+p+v} B(\xi, \eta)$$

where

$$\mu_A(i, j) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i=u} \sum_{\eta=j-v}^{j=v} A(\xi, \eta)$$

are the window means of images A and B, respectively, and where

$$\sigma_A^2(i, j) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \cdot \{[A(\xi, \eta)]^2 - [\mu_A(i, j)]^2\}$$

and

$$\sigma_B^2(i, j+p) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j+p-v}^{j+d+v} \cdot \{[B(\xi, \eta)]^2 - [\mu_B(i, j)]^2\}$$

are the corresponding window variances.

The correlation shift p is along the axis in the direction of camera displacement. The maximum $\phi(p)$ occurs at the sought disparity p^*, linking the corresponding pair of points of the two images, one in each.

The window size can be adjusted and is an important factor in the analysis. If too small a window is used, false matches can occur through random effects, the sample size being inadequate for reliable peak finding over the $\phi(p)$ function. Too large a correlation window leads to poor spatial discrimination of picture cells corresponding to different depths in the scene. In general $\phi(p)$ is a multimodal function which needs to be searched over using global optimization techniques [23], the simplest but most expensive of which is exhaustive search. Hierarchical searches including coarse and fine components can reduce the search cost considerably.

Yakimovsky and Cunningham [24] present details of a camera model and calibration system to support stereo disparity range finding. They continue with the description of stereo correlation algorithms in which specially configured masks are used instead of the more usual rectangular image window (see Fig. 7). This first type of mask [Fig. 7(b)] is a

set of concentric diamonds $D_0, D_1, \cdots, D_k$ where $D_0 = T_1$, the reference image point (I_1, J_1) and

$$D_i = \{I, J: \ |I_1 - I| + |J_1 - J| = d_i\}, \quad i = 1, \cdots, k.$$

Typical values of d_i are

$$d_1 = 1, \ d_2 = 2, \ d_3 = 4, \ d_4 = 8$$

giving an $N = 61$ point mask.

The second type of mask [Fig. 7(a)] consists of four line segments defined by integer k:

1) horizontal $(I_1 - k, J_1)$ to $(I_1 + k, J_1)$
2) vertical $(I, J_1 - k)$ to $(I, J_1 + k)$
3) 45° $(I_1 - k, J_1 + k)$ to $(I_1 + k, J_1 + k)$
4) -45° $(I_1 - k, J_1 + k)$ to $(I_1 + k, J_1 - k)$.

k is typically =8 producing an $N = 65$ point mask.

In both cases the mask definitions can be interpreted as sequences of N displacement pairs $(\Delta I_i, \Delta J_i)$, $i = 1, \cdots, N$, each point, m_i, on the mask being defined as $(I + \Delta I_i, J + \Delta J_i)$ with the mask centered on (I, J). To find T_2 in the second image of a stereo pair which represents the same scene point P which appears at T_1 in the reference image a mask correlation search is carried out along a line segment S. The search is indexed by k in the X direction.

The image intensity value X_i at each point m_k of the mask centered at T_1 on the reference image is stored in an N element array.

With the mask centered on $T_k = (I_k, J_k)$ in the second image, the intensity value Y_i at each point $m_i = (I_k + \Delta I_i, J_k + \Delta J_i)$ is sampled.

The correlation function

$$C_k = \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}) \Bigg/ \left(\sum_{i=1}^{N} (X_i - \overline{X})^2 \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \right)^{1/2}$$

$$-1 < C_k \leqslant 1$$

where

$$\overline{X} = \sum_{i=1}^{N} X_i / N \quad \text{and} \quad \overline{Y} = \sum_{i=1}^{N} Y_i / N$$

is transformed to

$$N \cdot \sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i \right) \Bigg/ \left\{ \left[N \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i \right)^2 \right] \cdot \left[N \sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i \right)^2 \right] \right\}^{1/2}$$

to minimize computation. Since the X_i's remain fixed while T_1 is fixed during the search for the corresponding point in the second image T_2, the maximization of the following suffices to determine the value of k which maximizes C_k:

$$(D \cdot |D|) \Bigg/ \left[N \cdot \sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i \right)^2 \right]$$

where

$$D = \sum_{i=1}^{N} Z_i Y_i - S_k \sum_{i=1}^{N} Y_i$$

$$S_k = \sum_{i=1}^{N} X_i$$

$$Z_i = N X_i \quad \text{for each } T_k.$$

Since S_k and Z_i are constant for the search over k, the main computational effort is in computing

$$\sum_{i=1}^{N} Y_i, \quad \sum_{i=1}^{N} Y_i^2, \quad \text{and} \quad \sum_{i=1}^{N} Z_i Y_i.$$

Two tests are applied to determine whether a particular mask at T_1 will produce reliable correlation peaks corresponding to a proper match. If the intensity variance of the mask points is less than 0.3 of the camera noise variance, the mask is not considered suitable for correlation matching, there being insufficient visual information present. If the autocorrelation function of T_1 centered mask points and mask point sets of a sequence of shifts along the X direction on the same image shows a sharp drop from 1 away from the reference position, the correlation with the other image proceeds. These measures are consistent with the "busyness" measures suggested earlier in this section.

Moravec [25] used a TV camera on a horizontal rack to gather a sequence of nine images at equal intervals of camera displacement. The high degree of redundancy this sequence provides was exploited to improve the accuracy of range estimation based on disparity. An experiment begins with a camera calibration phase with a visual chart which automatically establishes focal length and image distortion parameters. Then with the cameras pointing at the scene, localized features which can unambiguously be detected from different views are selected. Image regions with high contrasts in orthogonal directions (corners) are ideal. An "interest operator" subroutine attempts to select a scattering of such regions so that each object might be represented a few times. Sums of squares of adjacent pixel intensity difference in each of the directions, horizontal, vertical, left diagonal, and right diagonal are calculated over small square windows and the minimum of these four directional variance measures used as that window's "interest" measure. Points of interest valuable for disparity locally maximal interest points with other images by searching a whole image area or a specified rectangular subimage.

A hierarchical search which begins with a coarse strategy applied to a reduced resolution image and proceeds by refining the search into finer and finer resolution images guided by higher level results is used. The fifth image of the nine camera images sequence is used as the reference both for the interest operator phase and the correlation match phase. The correlator attempts to match selected high interest features from that image with each of the other eight images. Since the camera shift is horizontal, the search is restricted to narrow

horizontal bands. After the correlation search, each feature's position in each nine images is known. The 36 image pairings are used for stereo disparity analysis for each feature. The range estimations for each pair are considered to be the means of normal distributions with standard deviations inversely proportional to the relative camera shift, the distribution area being scaled by a confidence measure based on the correlation measures (the product of the two best match correlation figures with respect to the reference image using the value 1 for one factor when the central image is involved amongst the 36 pairs) and by the projection of the feature shift on the X axis. For each feature, the peak of the summed 36 distribution functions provides the overall range estimation with considerable reliability. False matches tend to produce distributions which fail to gain reinforcement from the others. In all, this method of combining the 36 estimations is statistically sound and most intelligent in refining the solution. Note that only eight correlation match searches for each feature of the reference image are involved.

Baker [26] describes a stereo pair range analysis technique based on edge data in the images. The use of edge data fulfills the basic requirement of visual "busyness" (at least in the direction across the edge) for reliable correlation matching, at the same time reducing the computational cost. Camera shift is in the direction of horizontal scan lines and only edges with a vertical component in their slope are used in the correlation process (i.e., edges are associated with sharp differences in an intensity plot along a horizontal scan line). The correspondence problem is attacked one horizontal line at a time using edge correlation procedures for finding the best association of first and second image edges; the information used is strictly local for this phase. At a more global level, edge continuity constraints are used to confirm or reject these edge pairings. This second phase is termed "cooperative continuity enforcement." Inconsistent pairings involving those edges where nearest image space connections (as seen in either image) are with edges other than given by the correlation based link, are removed. Those that remain hopefully provide reliable range data from the corresponding disparity values. This approach is an excellent example of filtering local information through a global constraints function to preserve consistency and thus improve reliability. This method should be particularly useful for colinear edged planar surfaced objects under edge enhancing lighting conditions.

The work by Marr and Poggio [27] has excited a considerable amount of interest among computer vision and psychophysicists alike, particularly as their proposals for solving the stereo disparity correspondence problem by the use of cooperative computational processes contains clues of human neurophysiological function in this same domain. Julesz's [28] findings regarding the human interpretation of random dot stereograms when viewed binocularly to yield patterns separated in depth suggests a mechanism of local processing which inspired Marr and Poggio in their computer vision work on stereo disparity analysis. A cooperative algorithm is one which operates in parallel upon a large array of inputs to yield a global (consistent) organization through local interacting constraints. In this case the constraints are derived from the physical 3-D world of solid objects where

1) a point on a surface has unique position in space at any time instant, and

2) the surfaces of objects are smooth compared to their range from the viewer and matter, divided into objects, is cohesive.

Only identifiable features on surfaces are suitable for matching stereoscopically; lines, edges, shadows, other markings, etc. in the images usually have a physical existence in the 3-D scene.

The above two constraints can be mapped into rules for combining descriptions (including positions) of identifiable features in the left and right images of a stereo pair.

1) *Uniqueness:* Each item (feature) can be assigned at most, only one disparity value.

2) *Continuity:* Disparity varies smoothly almost everywhere, i.e., discontinuities corresponding to depth change occur only relatively infrequently in the image when compared with the total area.

Computational cells for each x, y position in the image pair and for each possible disparity value d, evaluate the state for triples in x, y, d to represent actual disparity match points by using iterative processes with the local neighborhood constraint conditions inhibiting and supporting candidature at each step. The stable states for the cells represent a disparity solution. The computational cost, when the algorithm is processed on a conventional serial machine would probably be large, but specialized array processing would be most effective in reducing this cost.

The form of the iterative equations is

$$C_{xyd}^{(n+1)} = \sigma\left\{\left(\sum_{x'y'd' \in S(x,y,d)} C_{x'y'd'}^{(n)}\right) - \xi\left(\sum_{x'y'd' \in O(x,y,d)} C_{x'y'd'}^{(n)}\right) + C_{xyd}^{(0)}\right\}$$

where

1) $C_{xyd}^{(n)}$ is the state of the cell at position (x, y) with disparity d at iteration n;

2) S and O identify supportive neighbors and inhibitive neighbors, respectively, in the vicinity of x, y, d;

3) σ is a sigmoid function ("S" shaped curve) with range $[0, 1]$;

4) ξ is the inhibition constant.

From the paper, it would seem that when σ is a simple threshold function, the process converges for a wide range of parameter values. A number of difficulties are encountered in regarding this process as a theory of human vision stereo. These concern the human tolerance for the defocusing of one image, the movements of the eyes as a stereo pair of images come into fusion and the hysterisis effect by which there is a delay in matching but fusion remains for subsequent separation of the images in the pair beyond the distance for which fusion was initially impossible.

More recently, Marr *et al.* [29]-[31] have proposed an alternative theory of stereo vision computation which has strong links with low-level biological visual mechanisms. It is based on initially extracting edges with mask operators of various sizes convolved over both left and right images and extracting the zero crossings for each. The stereo correspondence problem is then solved by using the disparity matches of the gross line structures for the results of using the large masks to guide the matchings at finer, higher resolution.

Neurophysiological studies carried out on cats and monkeys indicate a lateral inhibition local operator which can be modeled mathematically as the difference of two Gaussian distributions:

$$G_1(x,y) - G_2(x,y) = \frac{1}{2\Pi\sigma_1} e^{2\sigma_1^2/-r^2} - \frac{1}{2\Pi\sigma_2} e^{2\sigma_2^2/-r^2}$$

where r is radius from the center at the point of reference (x, y) and σ_1, σ_2 are standard deviations which correspond to scale factors for excitatory (G_1) and inhibitory (G_2) distributions, respectively. This is approximately equivalent to the application of a Gaussian smoothing operator followed by the application of the Laplacian operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

which is a nonoriented second derivative.

Since convolving the image I with G, the Gaussian smoothing mask, and then applying ∇^2 is equivalent to applying, in one pass, the convolution of the product mask $\nabla^2 G$, the computational cost of convolutions over finely quantized images is reduced:

$$\nabla^2(G * I) = (\nabla^2 G) * I$$

where $*$ is the convolution operator.

The shape of the $\nabla^2 G$ mask is given by

$$\left(2 - \frac{r^2}{\sigma^2}\right) e^{-r^2/2\sigma^2}.$$

The various scales of edges are extracted by applying the discretized form of this function, with a geometric progression of sizes (equivalent to adjusting σ) and extracting the zero crossings (which correspond to extrema of the first derivatives of the smoothed images).

Intermediate between the extraction of zero crossings for a sequence of scaled $\nabla^2 G$ mask convolutions and application of stereo correspondence algorithms, a representation called the "raw primal sketch" [32] is created by segmenting collections of zero-crossing contours into sequences of short line segments and evaluating, for each, its position, orientation, length, and rate at which $\nabla^2 G$ changes across the segment. This representation aids the left/right image matching process which, as mentioned earlier, is applied at the crudest scale level first, these results then being used to guide finer matches at higher resolution.

It is given in [33] that the first implementation of the Marr-Hildreth theory took in the order of 3 h to compute the coarse level zero crossings of a 512 X 512 pixel image and a prototype hardware implementation some 30 min. In [34] is a report of a hardware implementation which can complete the zero crossings in under 0.25 s. Note, however, that the smallest operator used on something like a 512 X 512 pixel image (N^2) is 35 X 35 pixels (M^2) and a convolution pass involves some M^2N^2 multiplications and slightly fewer additions.

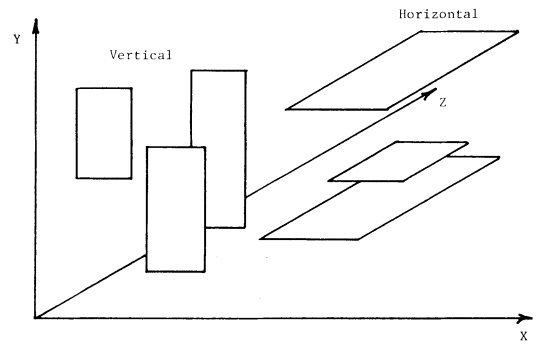


Fig. 8. Horizontal and vertical surface assumption.

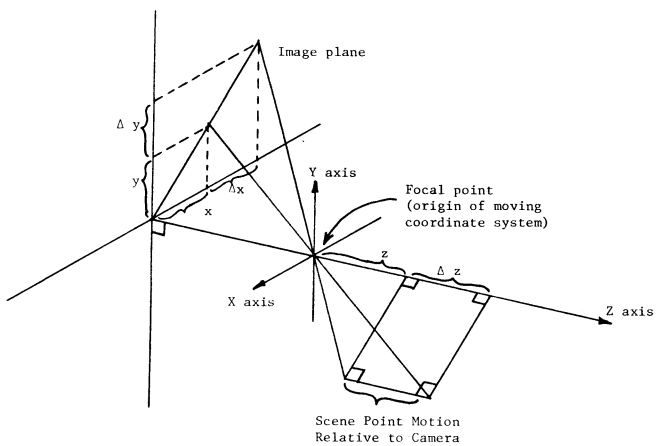


Fig. 9. Relationship between distance in the scene and image displacement.

VIII. Range from Camera Motion

In this case camera motion is not restricted to a limited lateral displacement as for stereo disparity evaluation. Two approaches in this class are represented by Williams [35] and Prazdny [36].

Williams makes the simplifying assumption that all surfaces are planar and orientated in one of only two directions, vertical and horizontal (see Fig. 8). The relationship between the camera relative movement of a point in the scene and the corresponding displacement in the image is illustrated in Fig. 9, where from similar triangles

$$\Delta x = \frac{x\Delta z}{z} \quad \text{and} \quad \Delta y = \frac{y\Delta z}{z}$$

where z is the distance to the scene point at time t_1, Δz is the camera movement since t_0, (x, y) are the coordinates of the corresponding image point at t_0, and $(\Delta x, \Delta y)$ are its displacement components at t_1. The relative distance between camera and scene has diminished between t_0 and t_1. All points in the image move radially outward from a point on the image plane called the focus of expansion. It is assumed that the position of this point is known and that there is no movement in the scene itself. An initial static segmentation is used in defining the extent of the planar regions.

A simple model of planar surfaces (either vertical or horizontal), used to express a 3-D scene interpretation, is used to

predict image dynamics which are tested against image data by applying the prediction (in reverse time) to an image at t_1 and calculating an error function based on differencing (pixel by pixel) for each surface region, this predicted (synthetic) image with the previous time sequence actual image at t_0.

The averaged difference for each region is that surface's error value. Subpixel displacement resolution is achieved by interpolation based on weighting the gray levels of each pixel in the predicted window by the areas of the pixel cells involved.

Occlusion predictions are used to prevent companions of region parts not visible in both time sequence images of a pair. A search process to reduce the error measure refines the scene interpretation model. The search is split into two independent parts, each involving only one parameter per surface. The first search is to find the distance Z for each surface assuming it is vertical and the other to find the height Y for each surface assuming it is horizontal.

The distance to all surfaces under the vertical orientation assumption are refined to reduce the appropriate error value. Simultaneously and independently the heights of each surface are refined by the second search on the assumption of horizontal orientation. The correct orientation for each surface is decided on the basis of lowest error. Unresolved errors in the initial static segmentation can be detected once the surface model is refined.

Each synthetic image which is tested against a real image reflects a set of systematic changes in the distance and height of every surface—these are simply increments and decrements of Z and Y for each hypothesized surface. The global minimum error for each surface is sought using a hill climbing algorithm (see [23] for survey on global search methods) with fractional perturbations on the best values of Zs and Ys found so far. The fractional perturbations are diminished as extrema are approached and a number of simple stopping criteria are applied for each surface independently.

The overall approach is a conventional optimization strategy applied independently for each surface. Since the error evaluations are based on average pixel value differences over segmentation regions for each of the Z and Y values for the corresponding surface the process is computationally expensive, particularly if a large number of surfaces are involved. Again, accuracy would diminish for distant objects as corresponding image displacements with camera motion would be small. Overall, this method represents an interesting approach worthy of further investigation but would seem to be both tedious and not particularly reliable, especially since the initial static segmentation is carried out without the aid of range information. Also, the sharpness with which a synthetic to real image segment segment match can be achieved would depend upon the visual "busyness" in those regions. Furthermore, hill climbing techniques are intrinsically only able to find local extrema which are obviously not guaranteed to be the global ones sought if multimodal search performance index functions are involved.

In [35] Prazdny presents a rather elegant method of recovering instantaneous egomotion (observer motion) parameters and a surface normal map (from which relative range information can be derived) starting with optical (or retinal) flow data in the form of the instantaneous positional velocity field (planar retina based) which is regarded as being provided by some procedure not yet defined. The vector geometry is complex but elegant and the method involves an iterative solution of a set of three third degree equations in three unknowns for retinal point sets. The author admits that the greatest weakness in the approach is the assumption concerning the provision of the instantaneous positional velocity field. Apart from the assumption concerning the provision of this field defined on the observer's retinal plane, the only requirements are the smoothness of the observer trajectory and the rigidity of the objects in the scene. However, absolute distance to an object is not recoverable and must be seen as another weakness of the method. The computational complexity would also be considerable.

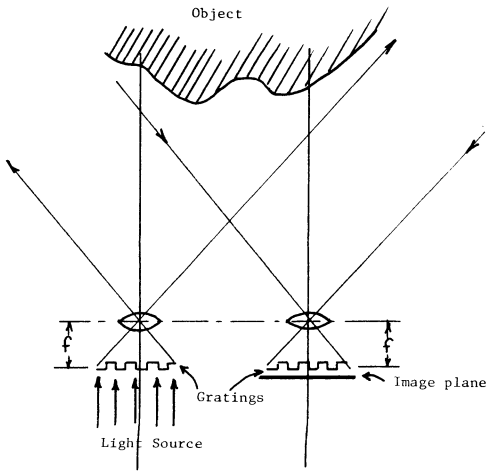


Fig. 10. Moiré fringe range contour apparatus.

IX. Moiré Fringe Range Contours

A Moiré fringe interference pattern formed by illuminating a scene with shadow patterns through an equispaced optical grating and viewing the scene through an identical grating in a camera displaced laterally from the light projection system (see Fig. 10) represents contours of equal range, but the sign information indicating increasing or decreasing range between adjacent contour lines is missing. Completing two experiments with a known movement of the scene objects between observations or using a phase shifted second grating does allow sign recovery, but contour correspondence problems make range recovery difficult.

Idesawa *et al.* [37], [38] describe an ingeneous method whereby automatic range recovery is possible by modifying the standard Moiré fringe method through use of a high spatial resolution image tube. In this method, the second grating is replaced by a "virtual grating" formed by a set of equispaced scan lines of the image tube system. Sampling along these lines is equivalent to the superposition associated with the standard configuration of Fig. 10. Suppression of unwanted lines not associated with the intensity peaks and valleys which form the range contours is carried out to clarify the contour patterns.

Contour lines for different range levels can be produced simply by changing the phase or the pitch (spacing) of the scanning lines, while using just one grating shadow lit image.

The required contour change sign information is recoverable in this process (using only phase shift will suffice). The image tube spatial accuracy and reproducibility was an order of

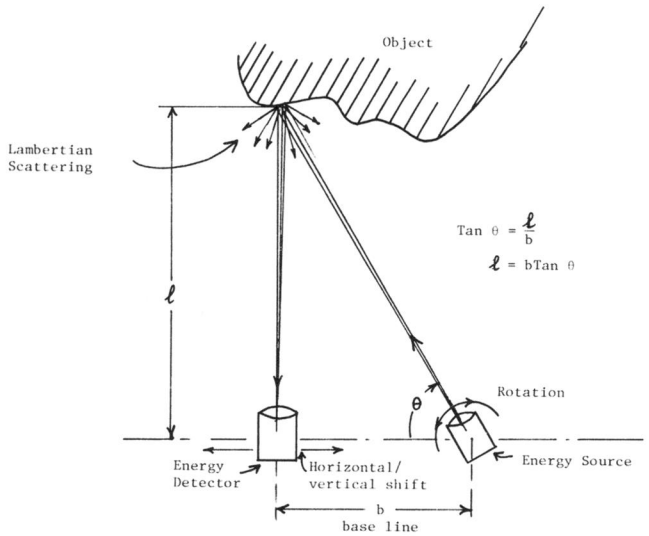


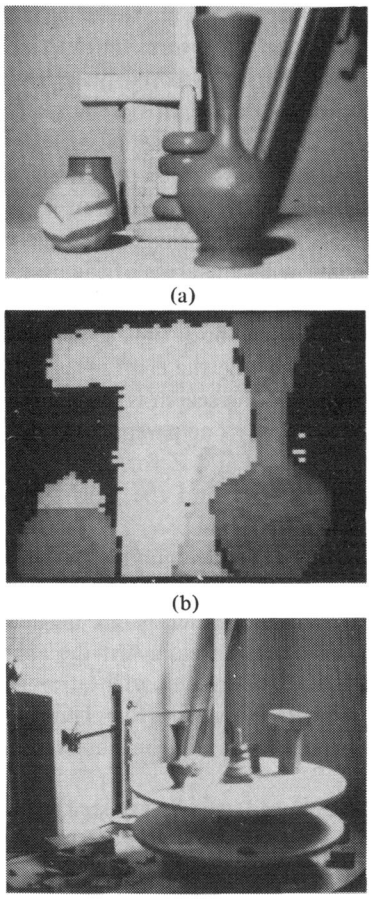
Fig. 11. Simple triangulation range finding geometry.

magnitude higher than for an ordinary cathode ray tube. Unfortunately, the experiment required the photographic recording of the grating lit image and this was scanned in a flying spot scanner mode using the high resolution image tube as the spot light source. If high resolution solid-state or vidicon TV cameras could be used directly it would be a great advantage for real time range analysis of scenes which may be required to be robotically navigated through or manipulated. The basic concept is certainly worth exploring further as the potentials are high. However, once again, as for many of the ranging methods presented earlier, although relative ranges over contiguous surfaces can be measured this way, the absolute range to a partially occluded surface cannot be recovered if there is no range contour continuum to that surface.

X. Simple Triangulation Range Finder

Perhaps the most obvious method of absolute range finding is to use simple one spot at a time triangulation. In a sense, this is a one-dimensional version of stripe light ranging and no image analysis is required. The image of a small portion of the scene is focused upon a light detector. A narrow beam of light from a source laterally displaced from the detector is swept over the scene. The known directions associated with source and detector orientation at the instant the detector "sees" the light spot on the scene are sufficient to recover range if the displacement between the source and detector is known. It is sensible, of course, to sweep the light beam only in the plane defined by the line from the scene to the detector and the line from the light source to the detector. If the detecting system is made to "look" at a raster sequence of scene points, sweeping the source beam in the suitable plane for each position and recording the relevant angles when a "strike" is detected, a reliable rangepic can be easily constructed (see Fig. 11). Sometimes no strike will be detected because of occlusion or surface absorbance. The larger the base line distance between detector and source, the more accurate the ranging but more prevalent the "missing parts" problem caused by directional occlusion. Also, closer ranges can be more accurately measured.

Fig. 12(b) shows a 64 X 64 rangepic of the scene of Fig. 12(a) obtained by using the infrared range detection components from an inexpensive Cannon AF 35 mm camera and mounting them on the pen carriage of an XY plotter, which was driven in a raster sequence. The experimental setup is shown in Fig. 12(c). The scan time was in the vicinity of 50 min but there is no inherent reason why the whole process could not be speeded up by an order of magnitude or two with the design of suitable apparatus. Range accuracy for the above example was not great but this could easily be improved upon also. This range scanner is also discussed in [39]. The use of an infrared source permitted range finding in normal lighting conditions (or in the dark).



(a) (b) (c)

Fig. 12. (a) Color image. (b) Pseudocolor rangepic of scene depicted in (a). (c) Infrared range scanner experimental setup.

XI. Time-of-Flight Range Finders

A distinct turning away from triangulation based range finding with its inherent "missing parts" problems and diminishing accuracy with range is exemplified in the existence of time-of-flight ranging apparatus where energy source and detection windows can be coaxial and range accuracy maintained over depth up to the point where reliable signal detection is no longer possible. The main two representatives in this category are ultrasonic range finders and laser range finders, the speed of sound and the speed of light, respectively, being the most relevant parameters. No image analysis is involved, nor are assumptions concerning the planar properties or otherwise of the objects in the scene relevant. Furthermore, absolute range is directly available and rangepic registration with imagery easily achieved. Since the range measurements are image independent, they are a legitimate source of com-

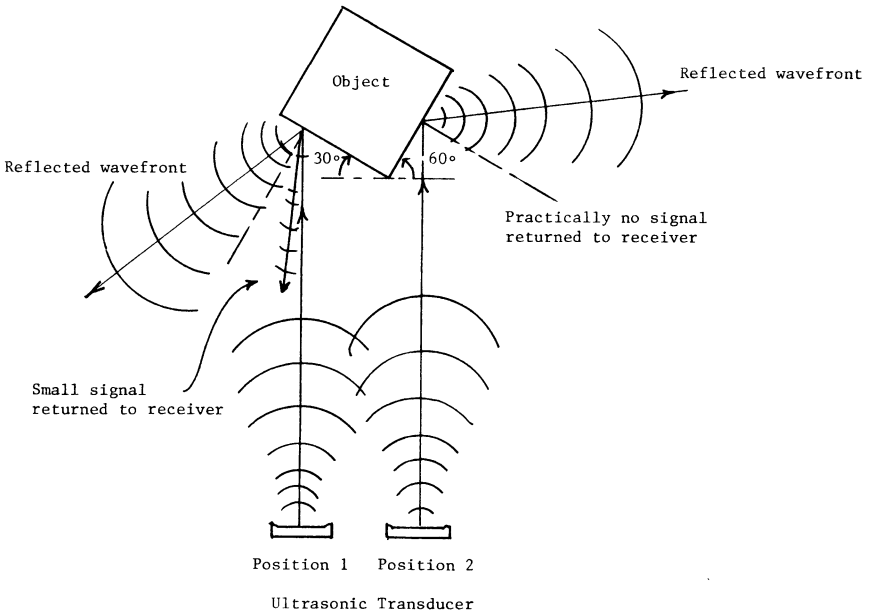


Fig. 13. Ultrasonic signal reflection.

plementary information that can support reliable scene segmentation, unrestricted by domain specificity.

A whole range of anthropomorphically phrased questions are entirely sidestepped as direct time-of-flight ranging has no analog in human depth perception, although, of course, it is relevant to bat navigation.

A. Ultrasonic Range Finding

Polaroid makes available an ultrasonic ranging system kit based on the transducer and electronics of their ultrasonic range finder cameras. This kit, which includes a test board and range read-out display and a fairly comprehensive manual, is ideal for exploring the advantages and disadvantages of ultrasonic rangefinding for any particular application.

A 1 ms ultrasonic "chirp" consisting of 56 pulses at four frequencies, 60, 57, 53, and 50 kHz, is transmitted by a simple electrostatic transducer of about $1\frac{1}{2}$ in diameter and exhibiting a beam pattern with a major forward lobe of about 30° solid angle. The signal reflected off an object within range is detected by the same transducer and processed by an amplifier whose gain and bandwidth are adjusted during delay between "chirp" and "bounce back" so as to improve the signal noise ratio and the reliability of signal detection. The mixed frequency "chirp" is used to lower the probability of signal cancellation for certain target topographies. Ranges from 0.9 to 35 ft are recovered with an accuracy of about 1 in. Two basic problems are encountered in using such a device in an attempt to derive a rangepic of acceptable spatial resolution for computer vision studies perhaps involving robotic manipulation. The first is that the 30° solid angle of the main lobe of the transducers beam pattern does not allow better than about 4 X 4 resolutions of a 90° solid angle field. Special acoustic focusing devices could improve this resolution as could using arrays to sharpen the directionality; a simple sound absorbing plastic foam tube is also effective in narrowing the directionality to about a 10° solid angle but even this only gives at best a 10 X 10 resolution over a 90° solid angle field [40]. The second problem is a more fundamental one and concerns the intrinsic properties of acoustic waves and reflecting surfaces. If the transducer disk is pointed at more than about 40° to the normal of a large hard surface there is a tendency for the acoustic wave to bounce off mostly concentrated in a direction where the angle of incidence and the angle of reflection are equal (see Fig. 13); consequently, little energy is reflected directly back to the detector directly from this surface. Sometimes other objects in the path of the reflected beam may return signals back to the sensor via reflection off the plane, thus producing a false reading. No return energy is a better result since it would indicate an invalid situation for the range finder. This reflection effect is explained in terms of Huygen's principal and the undulations of the surface material in relation to the wavelength of the energy. A simple particle theory analogy is the way in which many elastic balls whose sizes (diameters, say) correspond to energy wavelength would bounce off a relatively smooth surface (whose undulations are small in comparison with the ball's diameter) in a highly predictable way with a small probability of returning towards the incident direction. When the surface is relatively undulating in comparison with the ball size a stream of balls striking a small portion of this surface would bounce off in all directions with equal probability; some energy is detectable along the incident direction. For light sources the surface needs to be almost mirror smooth before this specular reflectance effect is noticeable since the wavelengths involved are much smaller than in the ultrasonic range. Thus for a large number of commonly encountered surface materials the ultrasonic device cannot measure range at incidence angles more than about 40°. The scattering of reflected energy with equal probability in all directions in a hemisphere on the surface is known as lambertian scattering in the theory of light.

In summary, range finding using a system like the Polaroid range finder kit is not suitable for producing medium to high resolution rangepics over scenes containing hard objects with surfaces whose normals are in arbitrary directions. However, for crude navigation purposes, the device would be most useful as an obstacle detector.

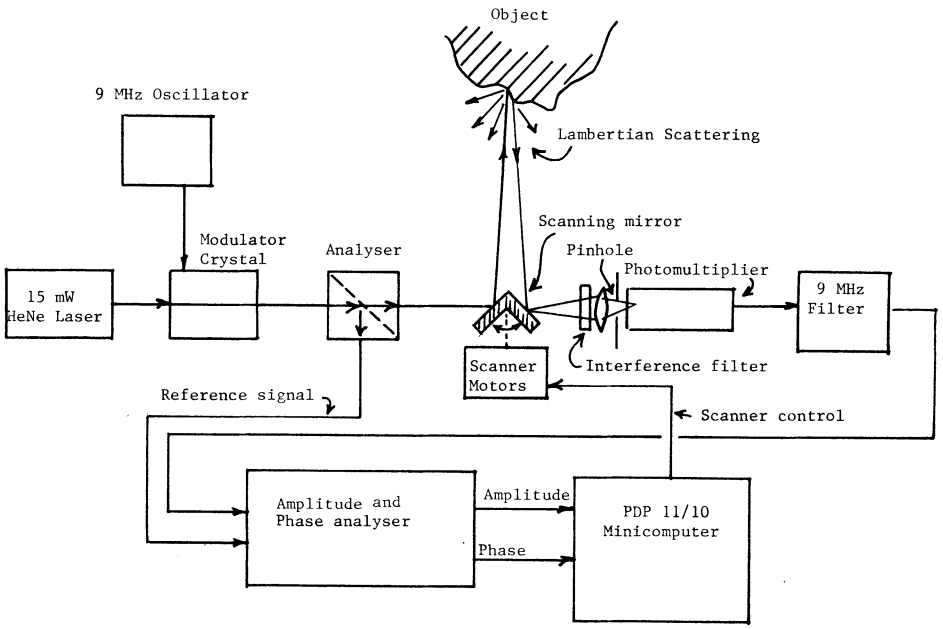


Fig. 14. Phase detection laser range finder block diagram.

B. Laser Range Finders

There are two basic laser range finder designs dependent upon time of flight to and from an object point whose range is sought. The first kind measures phase shift in a continuous wave modulated laser beam between leaving the source and returning to the detector coaxially. The second measures the time a laser pulse takes to go from the source, bounce off a target point (approximately lambertian surface assumed) and return coaxially to a detector. As light travels at approximately 1 ft/ns, the supporting instrumentation must be capable of 50 ps time resolution for range accuracy in the vicinity of $\pm\frac{1}{4}$ in. For both approaches an intrinsically large dynamic range of return energy is involved both because of the inverse fourth power range law involved and because of the variable reflectance properties of the target surfaces.

The modulated beam phase shift measuring version is represented by the instrument built at the Stanford Research Institute, reported by Duda and Nitzan [41] and detailed by Nitzan *et al.* in [42]. A simplified block diagram of this instrument is shown in Fig. 14. A scanning mirror unit points the modulated continuous laser beam at a raster scan of positions in a scene and captures a coaxial portion of the lambertian scattered beam for a receiver chain consisting of an interference filter, photomultiplier, logarithmic amplifier, and phase detector, the last using a sample of the direct source beam as a phase reference. Range is recovered through phase shift measurements and reflected intensity by energy measurements. The ratio of returned energy over source energy, when corrected for range gives the intrinsic surface property of the target known as albedo which is independent of both the surface orientation and the illumination. The combination of intensity and range information is a powerful complementary source of information for supporting scene segmentation and other scene analysis problems. The coaxial paths of source and reflected beams ensure not only that no shadows are cast on the scene by any one object or edge on another surface but also that there are no parts of the scene which can be illuminated by the source but not "seen" by the detector. The "missing parts" problem, which is prevalent in all triangulation based ranging including stereo disparity, is entirely absent. A 15 mW He-Ne laser ($\lambda = 632.8$ nm) was used and a 9 mHz modulation applied. The wide dynamic range ($\approx$100 dB) of the reflected energy and low energy laser used made it necessary to integrate over many measurements for each position to reduce the uncertainty of measurement to an acceptable degree. This proved time consuming as a typical 128 X 128 rangepic of 7-8 bits accuracy in the 1-5 m range took 2 h. If further developments in technique and instrumentation can reduce this time by three orders of magnitude, a device most useful for near real-time robotic hand/eye coordination tasks would result. Adding color discrimination would also be valuable.

The direct time of flight pulse laser range finder alternative is represented by the instrument developed at CALTECH's Jet Propulsion Laboratory and reported by Lewis and Johnston [43]. The block diagram is shown in Fig. 15.

A solid-state gallium arsenide pulse laser emitting at wavelength $\lambda = 840$ nm was used as the energy source and a photomultiplier with a gallium arsenide photosurface with spectral sensitivity to match as the detector. Again, as for the phase shift type system, a mirror scanning system was used to deflect the beam and to collect a coaxial component of the reflected beam. The output from the detector is passed onto a chain of sensitive instrumentation normally associated with nuclear physics experiments. A constant fraction discriminator produces a time pulse corresponding to a point on the input signal at a constant fraction of the peak; this ensures stable pulse arrival timing independent of intensity which has a large dynamic range. The time between a reference time pulse produced at the moment of laser firing and the output of the constant fraction discriminator is converted into a relatively wide (2 μs) pulse whose height is proportional to the required time interval (time to pulse height converter). This height is averaged over many pulses and digitized for trans-

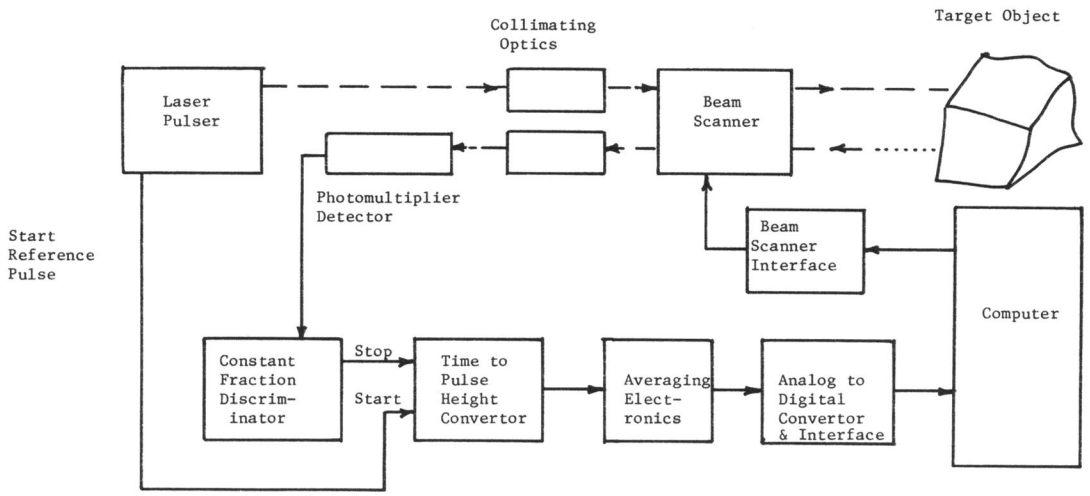
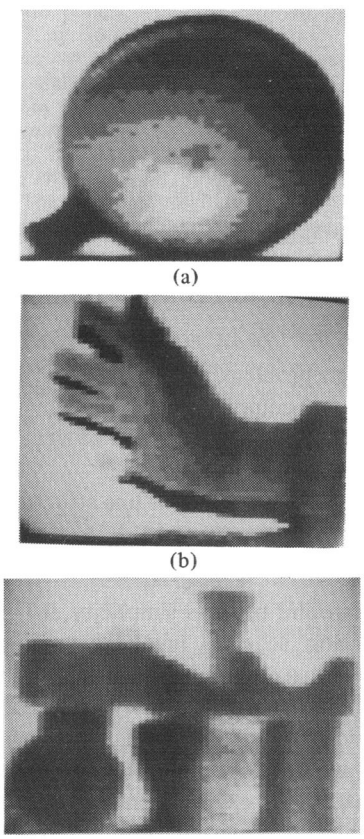


Fig. 15. Time-of-flight laser range finder schematic.

mission to the controlling computer. Its ranging accuracy is about 2 cm and once again the repetition requirements to increase signal/noise to acceptable levels constrained the speed with which the rangepic could be produced. The authors state that reliable range collection at beyond the rate of 100 points/s would be prohibited by the basic noise in the system. A range accuracy approaching 2 cm in the 1-3 m range was achieved. A 128 X 128 rangepic would, under ideal conditions take about 3 min to collect, still a long time in terms of a convenient vision-robotic manipulation or navigation cycle.

More recently Jarvis [44] has constructed a laser rangefinder using the same configuration as the Lewis and Johnston [43] instrument. Using a low powered infrared laser (820 nm) with a 100 ps pulse repeatable at 10 kHz, this instrument is capable of acquiring a noisy 64 X 64 rangepic in 4 s; a rather better result is achieved in 40 s with a resolution in the range 1-4 m of about $\frac{1}{2}$ cm. Example rangepics are shown in Fig. 16. Better accuracy in a shorter time can be achieved by increasing the laser power.

The laser ranging instruments described tend to be expensive to construct ($10 000-25 000) but with speed improvements they represent an effective direct attack upon the ranging problem with a wide variety of applications both on the laboratory bench and out of doors. Computational cost is minimal, all the complexity being delegated to a specialized piece of optoelectronic hardware.



(a) (b) (c)

Fig. 16. Examples of 64 X 64 rangepics from a laser time-of-flight range scanner (darker is closer). (a) Six inch diameter plastic funnel. (b) Human hand. (c) Block scheme.

C. Streak Camera Range Finders

Streak cameras (temporal dispersers) [45] accelerate photoelectrons from a photocathode (upon which incident light falls through a slit aperture) towards a positively charged mesh, and on passing through it, they are deflected in one direction (transverse to the acceleration) by an electrostatic field swept at variable speeds, the sweep being triggered by some timing reference (see Fig. 17). The intensity variation across this "streak" in the direction of the deflecting sweep gives the temporal intensity profile of the incident light at a time scale given by the sweep velocity. These cameras are capable of 10 ps resolution and are therefore suitable for discriminating light transit time variations corresponding to differences of light path distances of the order of 0.1 in (light travels at ≈1 ft/ns). Thus light energy arrival time variation due to reflection of a short duration laser pulse from objects at various distances from the temporal disperser camera can be used for time of flight range estimation. If a cylindrical lens is used to illuminate a scene with a line of pulsed laser light, it should be possible to obtain range information for one such line at a time. The light "sheet" from the laser can be deflected transversely by a scanning mirror system to give full coverage of the scene.

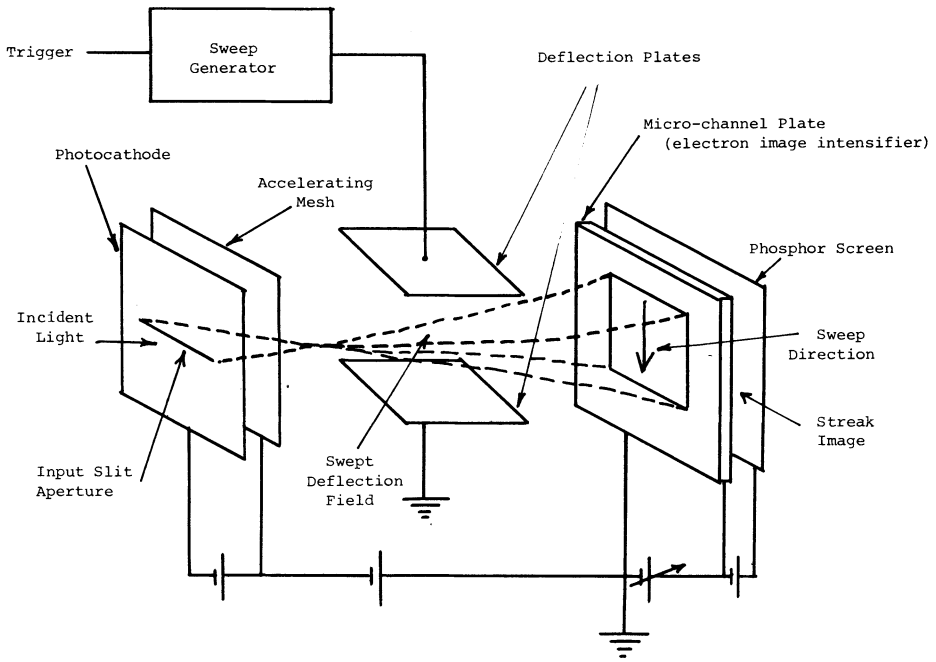


Fig. 17. Streak camera.

XII. Conclusion

Of the various approaches to range finding covered in this paper no one method would seem clearly superior to the rest. All appear to have drawbacks which fall into one or more of the following categories:

1) missing parts problem,
2) computational complexity,
3) time-consuming in improvement of signal/noise ratio,
4) limited to indoor application,
5) limited to highly textured or line structured scenes,
6) limited surface orientation,
7) limited spatial resolution.

From the viewpoint of sheer simplicity, it is hard to improve upon triangulation schemes involving one point at a time. In terms of potential, it would seem that direct laser time-of-flight range finders could, in theory, eliminate all the above problems provided an intense enough energy source could be provided—this would be at considerable expense and could also create a hazardous environment for humans. In terms of anthropomorphically posed questions, stereo disparity, occlusion, photometric and texture gradient methods would prove more interesting. From the practical standpoint of vision driven robotics, however, these approaches would hardly seem worthwhile as they are all, to some extent, indirect.

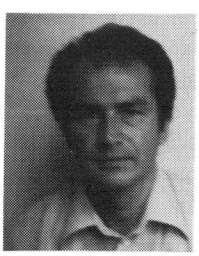
What is quite clear is that capturing the third dimension through nonimage-based range finding is of great utility in 3-D scene analysis since many of the ambiguities of interpretation arising from occasional lack of correspondence between object boundaries and inhomogeneities of intensity, texture, and color can be thus trivially resolved [46], [47].

For example, two objects of the same color and texture but at different ranges, which are in visual juxtaposition from the camera viewpoint may be difficult to separate through image analysis alone but can easily be detected as separate through rangepic analysis. An interesting approach to 3-D scene analysis using registered range and intensity data is given in [48]; in this case range analysis dominates the method and intensity data are used only when necessary. Collecting range data independently of intensity imagery analysis strengthens its ambiguity resolving potential. In both a technical and literal way, range data of this sort is orthogonal to the other data sources.

References

[1] R. N. Haber and M. Hershenson, *The Psychology of Visual Perception.* Holt, Rinehart and Winston, 1973.
[2] R. L. Gregory, *The Intelligent Eye.* New York: McGraw-Hill, 1970.
[3] M. L. Braunstein, *Depth Perception Through Motion.* New York: Academic, 1976.
[4] Y. Shirai and M. Suwa, "Recognition of polyhedrons with a range finder," in *Proc. 2nd Int. Joint Conf. Artificial Intell.*, London, Sept. 1971, pp. 80-87.
[5] G. J. Agin and T. O. Binford, "Computer description of curved objects," in *Proc. Int. Joint Conf. Artificial Intell.*, Stanford Univ., Aug. 20-23, 1973, pp. 629-640.
[6] T. O. Binford, "Visual perception by computer," in *Proc. IEEE Conf. Syst. Contr.*, Miami, FL, Dec. 1971.
[7] R. J. Popplestone, C. M. Brown, A. P. Ambler, and G. F. Crawford, "Forming models of plane-and-cylinder faceted bodies from light stripes," in *Proc. 4th Int. Joint Conf. Artificial Intell.*, 1975, pp. 664-668.
[8] G. F. Crawford, "The stripe finder hardware," Dep. Artificial Intell., Univ. Edinburgh, 1974.
[9] F. Röcker and A. Kiessling, "Methods for analysing three dimensional scenes," in *Proc. 4th Int. Joint Conf. Artificial Intell.*, 1975, pp. 669-673.
[10] P. M. Will and K. S. Pennington, "Grid coding: A preprocessing technique for robot and machine vision," in *Proc. 2nd Int. Joint Conf. Artificial Intell.*, Sept. 1971, pp. 66-68.
[11] D. Rosenberg, M. D. Levine, and S. W. Zucker, "Computing relative depth relationships from occlusion cues," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, Kyoto, Japan, Nov. 7-10, 1978, pp. 765-769.
[12] A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene labeling

by relaxation operations," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 420-443, 1976.
[13] S. W. Zucker, A. Rosenfeld, and L. S. Davis, "General purpose models: Expectations about the unexpected," in *Proc. 4th Int. Joint Conf. Artificial Intell.*, Tbilisi, Sept. 3-8, 1975, pp. 716-721.
[14] J. J. Gibson, *The Senses Considered as Perceptual Systems.* Boston, MA: Houghton-Mifflin, 1966.
[15] R. Bajcsy and L. Lieberman, "Texture gradient as a depth cue," *Comput. Graphics Image Processing*, vol. 5, pp. 52-67, 1976.
[16] R. M. Haralick, K. Shanmugan, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 610-621, Nov. 1973.
[17] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 269-285, Apr. 1976.
[18] B. K. P. Horn, "Focussing," M.I.T., Project MAC, AI Memo. 160, May 1968.
[19] R. A. Jarvis, "Focus optimisation criteria for computer image processing," *Microscope*, vol. 24, pp. 163-180, 2nd quarter, 1976.
[20] B. K. P. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," M.I.T., Project MAC, MAC TR-79, Nov. 1970.
[21] K. Ikeuchi and B. K. P. Horn, "An application of the photometric stereo method," in *Proc. 6th Int. Joint Conf. Artificial Intell.*, Tokyo, Japan, 1979, pp. 413-415.
[22] M. D. Levine, D. A. O'Handley, and G. M. Yagi, "Computer determination of depth maps," *Comput. Graphics Image Processing*, vol. 2, pp. 134-150, 1973.
[23] R. A. Jarvis, "Optimisation in adaptive control: A selective survey," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, pp. 83-94, Jan. 1975.
[24] Y. Yakimovsky and R. Cunningham, "A system for extracting three-dimensional measurements from a stereo pair of TV cameras," *Comput. Graphics Image Processing*, vol. 7, pp. 195-210, 1978.
[25] H. P. Moravec, "Visual mapping by a robot rover," in *Proc. 6th Int. Joint Conf. Artificial Intell.*, 1979, pp. 598-620.
[26] H. H. Baker, "Edge based stereo correlation," in *Proc. ARPA Image Understanding Workshop*, Univ. Maryland, Apr. 1980.
[27] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," M.I.T., A.I. Lab., Memo. 364, June 1976.
[28] B. Julesz, "Binocular depth perception without familiarity cues," *Science*, vol. 145, pp. 356-362, 1964.
[29] D. Marr and T. Poggio, "Computational approaches to image understanding," M.I.T., A.I. Lab., see also *Proc. R. Soc. London B*, vol. 204, pp. 301-328, 1979.
[30] D. Marr and E. C. Hildreth, "Theory of edge detection," in *Proc. R. Soc. London B*, vol. 207, pp. 187-217, 1980.
[31] E. C. Hildreth, "Edge detection in man and machine," *Robotics Age*, pp. 8-14, Sept./Oct. 1981.
[32] D. Marr, "Early processing of visual information," *Phil. Trans. R. Soc. London B*, vol. 275, pp. 483-524, 1980.
[33] M. Brady, Computational approaches to image understanding," M.I.T., A.I. Lab., A.I. Memo. 653, Oct. 1981.
[34] H. K. Nishihara and N. C. Larson, "Toward a real time implementation of the Marr-Poggio stereo matcher," in *Proc. Image Understanding Workshop*, Lee Bauman, Ed., 1981.
[35] T. D. Williams, "Depth from camera motion in a real world scene," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 511-516, Nov. 1980.
[36] K. Prazdny, "Motion and structure from optical flow," in *Proc. 6th Int. Joint Conf. Artificial Intell.*, Tokyo, Japan, 1979, pp. 702-704.
[37] M. Idesawa, T. Yatagai, and T. Soma, "A method for automatic measurement of three-dimensional shape by new type of Moiré fringe topography," in *Proc. 3rd Int. Joint Conf. Artificial Intell.*, Coronada, CA, Nov. 8-11, 1976, pp. 708-712.
[38] —, "Scanning Moiré method and automatic measurement of 3D shapes," *Appl. Opt.*, vol. 16, pp. 2152-2162, Aug. 1977.
[39] R. A. Jarvis, "A computer vision and robotics laboratory," *IEEE Computer*, pp. 8-24, June 1982.
[40] —, "A mobile robot for computer vision research," in *Proc. 3rd Australian Comput. Sci. Conf.*, A.N.U., Canberra, A.C.T., Jan. 31-Feb. 1, 1980, pp. 39-51.
[41] R. O. Duda and D. Nitzan, "Low-level processing of registered intensity and range data," in *Proc. 3rd Int. Joint Conf. Artificial Intell.*, 1976.
[42] D. Nitzan, A. E. Brain, and R. O. Duda, "The measurement and use of registered reflectance and range data in scene analysis," *Proc. IEEE*, vol. 65, pp. 206-220, Feb. 1977.
[43] R. A. Lewis and A. R. Johnston, "A scanning laser rangefinder for a robotic vehicle," in *Proc. 5th Int. Joint Conf. Artificial Intell.*, 1977, pp. 762-768.
[44] R. A. Jarvis, "A laser time-of-flight range scanner for robotic vision," Australian Nat. Univ., Comput. Sci. Tech. Rep. TR-CS-81-10; also in preparation for publication in *IEEE Trans. Pattern Anal. Machine Intell.*
[45] Y. Tsuchiya, E. Inuzuka, Y. Suzui, and W. Yu, "Ultrafast streak camera," in *Proc. 13th Int. Congr. High Speed Photography and Photonics*, Tokyo, Japan, Aug. 20-25, 1978.
[46] R. A. Jarvis, "Expedient 3D robot colour vision," Australian Nat. Univ., Comput. Sci. Tech. Rep., 1982.
[47] —, "Vision driven robotics in a partially structured environment," Australian Nat. Univ., Comput. Sci. Tech. Rep. TR-CS-82-03, 1982.
[48] R. O. Duda, D. Nitzan, and P. Barrett, "Use of range and reflectance data to find planar surface regions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 259-271, July 1979.



R. A. Jarvis received the Ph.D. degree in electrical engineering from the University of Western Australia in 1968.

He is currently a reader in computer science at the Australian National University, Canberra, Australia, where he was Head of the Department of Computer Science from 1976 to 1979. He spent 1969, 1970, and 1977 as a Visiting Professor in Electrical Engineering at Purdue University, West Lafayette, IN. His current research interests include digital computing technology, pattern recognition, image processing, computer vision, and robotics.