Representations for Active Vision

Cornelia Fermiiller and Yiannis Aloimonos
Computer Vision Laboratory
("enter for Automation Research
Department of Computer Science and Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Abstract

As the field of Computational Vision matures, more efforts are devoted to vision systems that are active and need to interact with their environment in real time. A prerequisite for integrating Vision and Action is the development of a set of representations of the visual system's space-time, where space includes the system itself. Thus we are faced with the problem of studying the nature of appropriate representations and also with the computational task of acquiring them in a robust manner and in real time. Both of these problems are addressed in this paper from a computational point of view. In particular, we study representations needed by active visual systems in order to understand their self-motion and the structure of their environment.

The representations are of less metric information content than the ones traditionally used, including depth, surface normals, curvature and 3-D metric values for the parameters of rigid motion, etc.; but they are rich enough to allow the system to perform a large number of actions. These representations, indexed in image coordinates, are the direction of translation and the direction of rotation for the case of motion and a monotonic function of the depth value in the case of shape description. Their advantage comes from the fact that they can be computed from minimal and well-defined input (flow or disparity values along image gradients), as opposed to the traditional ones which require image correspondence or the utilization of assumptions about the environment.

1 Introduction: Active Vision Revisited

If Computer Vision was once limited to the study of mappings of a given set of visual data into representations on a more abstract level, it now has become clear that Image Understanding should also include the process of *selective acquisition of data in space and time*. This has led to a series of influential studies published under the headings of Active, Animate, Purposive, or Behavioral Vision. However, with a formal theory integrating perception and action still lacking, most studies have treated Active Vision [Aloimonos *et* a/., 1988; Bajcsy, 1988; Ballard and Brown, 1992] as an extension of the classical reconstruction theory, employing activities only as a means to regularize the classical ill-posed inverse problems, in order to recover a metric representation of space-time which is general-purpose and can be used for accomplishing any task. In other words, the concept, of selective acquisition and processing of data in space-time was understood only through the manipulation of the geometric and physical parameters of the sensory apparatus (focusing, fixation, self-motion, etc.).

An important point was missed: that an intelligent vision system exists in space-time (where space-time includes the system itself) and in order for the system to function properly, i.e., to act appropriately in a variety of situations, it should be able to develop robust descriptions of space-time. That is, it should be able to develop representations that would be adequate for accomplishing a set of tasks. But, classical geometry modeling the 2 1/2-D sketch has been inescapably with us all the time. Recovering the depth of surfaces in view or quantities that, are subject of Differential Geometry such as surface normals and curvatures seems to have been about the only goal of 3-D Computer Vision in the past 35 years. During the past few years, however, it has been repeatedly argued that this could be a misplaced goal, simply because it is too difficult to recover such complete descriptions of shape and space [Faugeras, 1992; Koenderink and van Doom, 1991].

In this paper we propose novel representations of an active observer's space-time and we show how they can be obtained in real time using pattern matching techniques in the spatiotemporal gradients of the image intensity function. The essence of these representations is that they are derived from well-defined input (no correspondence of features is required) and from global computations not affected by local errors of various sorts. In particular, we present new solutions for two basic problems in Vision, the one of estimating an observer's motion and the one of recovering a shape description for an active binocular system whose exact extrinsic parameters are not known. The next section describes these representations.

2 Retinotopic Motion and Shape Representations

Consider an observer moving in some environment. Although the organism as a whole might move in a nonrigid manner head, arms, legs and wings undergo different motions—the eyes move rigidly, i.e., as a sum of instantaneous translation and rotation. The points where the translational and rotational velocity vectors intersect the retina, called the FOE (focus of expansion), and the AOR (axis of rotation), are respectively the representation of egomotion. That is, estimation of a system's egomotion amounts to recognizing two locations on the retina.

Similarly, consider a binocular observer fixating on a point in space. The traditional representation used for the observer's extrapersonal space is the depth of features in the scene in view, or its derivatives such as shape, curvature, etc. We argue that the shape representations to be recovered for an active vision system are different, from the metric ones traditionally used in the sense that they cannot provide the 2 1/2-D sketch or functions of it. We call such representations "qualitative structures"; they basically amount to recovering a function of the depth of the scene, where partial information about this function is available. More specifically, if $z(x, y)$ is the function providing the depth z at, points (x,y) in the image plane, and if /,, i — 1,2,..., n is a set, of functions not necessarily exactly known, then $fi(z(x, y)), i$ = 1, . . ., n is a set of qualitative structures of the scene. For example, if / is a monotonic function in z with the property $z(x,y) < z(x',y') <=> f(z(x,y)) < f(z(x',y'))$, then / constitutes a qualitative representation that we call "ordinal structure," since knowledge about, it only allows to order the values of the scene's depth. Functions with different properties amount to qualitative representations of a different nature. In particular, in this paper we will show that an active binocular observer not aware of its exact extrinsic parameters capable of fixating at environmental points can recover an ordinal structure of the scene, without relying on the computation of exact correspondence of features in the two images. The concept, of ordinal depth computations has its origin in the relief transformations suggested already by the turn of the century psychologist, Helmholtz and recently taken up in the work of various perceptual studies [Koenderink and van Doorn, 1991; Todd and Reichel, 1989; Carding *et a*/., 1993]. However, in these studies either correspondence is assumed or special assumptions are made about the structure of the scene in view.

3 Motion and Stereo Fields and Patterns of Gradients

Most current motion estimation techniques require the computation of exact image motion (optic flow or correspondence of features). This however amounts to an ill-posed problem. On the basis of local information only the component of the optical flow perpendicular to edges, the so-called normal flow, is well-defined, although in many cases, it is possible to obtain additional flow information for areas (patches) in the image. Similarly, the computational analysis of binocular shape perception has been based on the two-dimensional disparity field, which is a special case of a motion field. In a limited area of the image around the fixation point the disparity measurements are small and thus they can be treated in a differential manner, just as optical flow. As in the case of motion, on the basis of local information only the component of the disparity vector perpendicular to edges, the so-called normal disparity, is well defined. The basic thesis in this paper is the following:

(a) For the case of the motion problem there exists a set of orientation fields on the retina which possess the property that measurements of motion available along these orientations have a global structure, i.e., they form simple patterns whose location and form encode the 3-D motion parameters. (To obtain an intuition see Figures le and d; 3a, b and c; and 4a, b and c.)

(b) For the case of the stereo problem, there exists a set of orientation fields with the property that measurements of normal disparity that happen to be along these orientations have an ordinal structure regarding depth, i.e., their values are sufficient for ordering the depth of the corresponding scene points (Figure 7a). Measurements in one orientation field allow the derivation of partial ordinal structure. Successive fixations allow the merging of these representations by filling in ordinal depth information from different fixations, thus building up a global ordinal depth map.

4 The Case of Motion: Preliminaries

To begin with, let us review the geometry of visual motion. To obtain a better intuition we first use a spherical retina. If the motion is a translation t, the motion field is along the great circles containing the vector t (Figure la), pointing away from the Focus of Expansion (FOE) and towards the Focus of Contraction (FOC). If the motion of the eye is a rotation of velocity w where the rotation axis cuts the sphere at points AOR (Axis of Rotation point) and -AOR, the motion field is along the circles resulting from the intersection of the sphere with planes perpendicular to the rotation axis (Figure lb). For general rigid motion the motion field on the sphere is the addition of a translational field and a rotational field. However, as input to the motion interpretation process we do not consider the motion field but the sign of the projection of motion vectors on appropriately chosen orientations.

4.1 Selection of Flow Orientations

Two classes of orientations are introduced which are defined with regard to an axis. Consider an axis s passing from the center of a spherical eye and cutting the sphere at points N and S. The unit vectors tangential to the great circles containing .s define a direction for every point on the retina (Figure lc). These orientations are called s-longitudinal. Similarly, the s-latitudinal orientations are defined as the unit vectors tangential to the circles resulting from the intersection of the sphere with planes perpendicular to the axis s (Figure ld). At each point the s-longitudinal and latitudinal vectors are

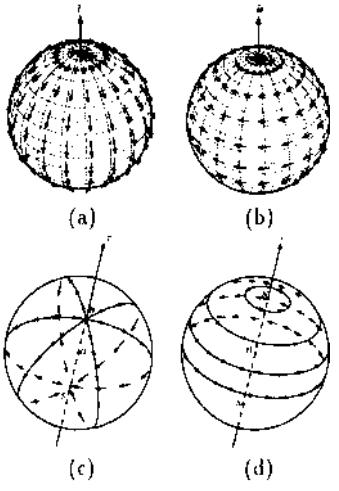


Figure 1: (a) Translational motion field, (b) Rotational motion field, (c) A longitudinal vector field defined by axis *s*. (d) A latitudinal vector field defined by axis *s*.

perpendicular to each other. Some properties of these directions will be of use later: Consider two axes s_1 (N_1 S_1) and s_2 (N_2S_2). Each axis defines at every point a longitudinal and a latitudinal direction. Figure 2 explains the locus of points where the s_1 longitudinal or latitudinal vectors are perpendicular to the .s_2 longitudinal or latitudinal vectors.

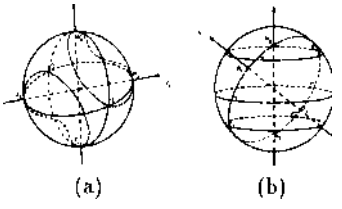


Figure 2: (a) On the sphere, the great circles containing S_1 and s_2 are perpendicular to each other on two closed second order curves, whose form depends on the angle between S_1 and s_2. These curves are defined as the set of points r on the sphere for which $(s_1 \times r) \cdot (s_2 \times r) = 0$ or $(s_1 r)(S_2 - r) = s_1 \cdot S_2$- (b) The S_1-longitudinal vectors are perpendicular to the s_2-latitudinal vectors along the great circle defined by S_1 and s_2.

Next, the structure of the projections of a rigid motion field on the S-longitudinal and latitudinal vectors is examined. More accurately, the sign of the projections of the motion field on the longitudinal and latitudinal vectors is investigated, since this is the information employed as input to the motion interpretation process. For this purpose it is necessary to agree upon a definition of the directions, *s (NS)* -longitudinal vectors are called positive (+), if they point away from *N,* negative (-) if they point away from 5, and zero (0) otherwise. Similarly, S-latitudinal vectors are referred to as positive (+) if their direction is counterclockwise with respect to S, negative (—) if their direction is clockwise, and zero (0) otherwise.

4.2 The Geometry of Image Motion Patterns

Since a rigid motion field is the addition of a translational and a rotational field, the cases of pure translation and pure rotation are first presented separately.

If the observer moves with a pure translation of velocity t, the motion field on the sphere is along the direction of the /-longitudinal vectors (Figure la). Projecting the translational motion field of Figure la on the S-longitudinal vectors of Figure lc, the resulting vectors will be either zero, positive or negative. The vectors will be zero on two curves as shown in Figure 2a (symmetric around the center of the sphere) whose shape depends on the angle between the vectors *t* and *s*. The area inside the curves will contain negative vectors and the area outside the curves will contain positive vectors (Figure 3a).

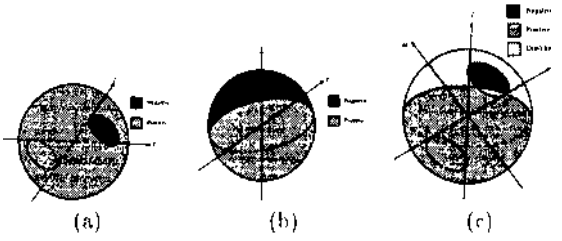


Figure 3: (a) Translational image motion along an *s*-longitudinal vector field: On two curves (like the ones in Figure 2a) passing through the points where /, -t, *s* and — *s* intersect the sphere the value is zero. Inside the curves the values are negative and outside they are positive, (b) Rotational image motion along an *s*-longitudinal vector field: On the great circle defined by w and *s the* values are zero. In one hemisphere the values are positive and in the other they are negative, (c) A general rigid image motion defines a pattern along every .s-longitudinal vector field: an area of negative values, an area of positive values and an area of values whose signs are unknown.

If the observer moves purely rotationally with velocity w;, the motion field on the sphere is along the direction of the w-latitudinal vectors (Figure lb). Projecting the rotational motion field of Figure lb on the *s (NS'*)-longitudinal vectors of Figure lc, the resulting vectors will be either zero, positive or negative. The projections will be zero on the great circle defined by *s* and w, positive in one hemisphere and negative in the other (Figure 3b).

If the observer translates and rotates with velocities t and w the projection of the general motion field on any set of s-longitudinal vectors can be classified for parts of the image. If at a longitudinal vector the projection of

both the translational and the rotational vectors is positive, then the projection of the image motion (the sum of the translational and rotational vectors) will also be positive. Similarly, if the projections of both the translational and rotational vectors on a longitudinal vector at a point are negative, the projection of the motion vector at this point will also be negative. Otherwise the sign will be unknown because it depends on the value of the translational and rotational vector components.

Thus, the distribution of the sign of image motion along the S-longitudinal set of directions defines a pattern on the sphere. Considering a general rigid motion field due to translation t and rotation w on an *s* (NS')-longitudinal set of directions, a pattern like the one shown in Figure 3c is obtained, which consists of an area of strictly positive values, an area of strictly negative values, and an area in which the values can not be determined without more information. The pattern is characterized by one great circle containing w and *s* and by two quadratic curves containing the points FOE, FOC, *N* and *S*.

It is worth stressing that the pattern of Figure 3c is independent of the scene in view and depends only on a subset of the 3-D motion parameters. In particular, the great circle is defined by one rotational parameter and the quadratic curve by two translational parameters. Thus the pattern is of dimension three. Also, the pattern is different for a different choice of the vector *s*.

Similarly, considering the projection of a rigid motion field on the *s* latitudinal directions (defined by the vector S*(NS)),* another pattern is obtained which is dual to the one of Figure 3c. This time the translational latitudinal flow is separated into positive and negative by a great circle and the rotational flow by two closed quadratic curves.

4.3 Egomotion Estimation Through Pattern Matching

The geometric analysis described above allows us to formulate the problem of egomotion estimation as a pattern recognition problem. Assume that the system has the capability of estimating the sign of the retinal motion along a set of directions defined by various S-longitudinal or latitudinal fields that happen to be perpendicular to the local edges. (In a number of cases the sign may be estimated in other directions as well). If the system can locate the patterns of Figure 3c in each longitudinal vector field (and the dual pattern in the latitudinal field), then it has effectively recognized the directions t and w. If information (positive, negative or zero) is not available in many directions, there might be an uncertainty in the computations in the sense that more than one set of patterns may be fitted to the data and the FOE and AOR may be obtained only within bounds.

For the case of a planar retina the latitudinal and longitudinal fields take a different form which is easily computed by projecting them on a plane tangential to the sphere (Figure 4a,b). In this case areas with negative and positive normal motion measurements lie in areas separated by a conic section (circle, ellipse, hyperbola, or parabola) and a straight line. Figure 4c pictures an example of a longitudinal pattern. Figure 5 shows results from experiments on planar images from an outdoor scene.

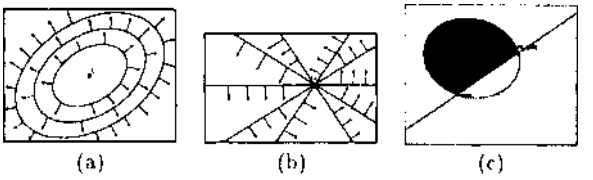


Figure 4: (a) In the plane the S-longitudinal vectors become perpendicular to conic sections defined by a family of cones with an axis parallel to *s*. (b) The S-latitudinal vectors become perpendicular to straight lines passing through the intersection s_0 of *s* with the plane, (c) In the plane the longitudinal vectors form patterns defined by a conic section and a straight line (dark negative, light positive, white "don't know"). Here, s_0 is denoted *as (fa/C,fB/C).*

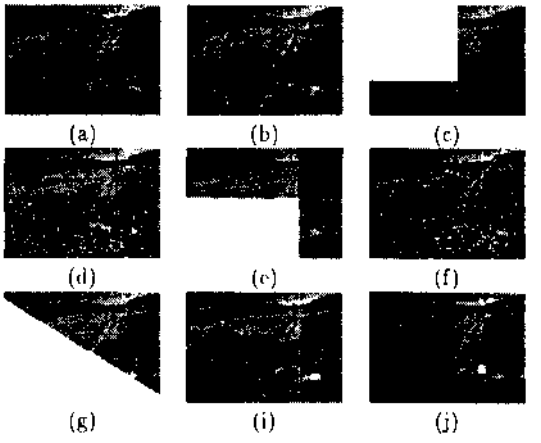


Figure 5: A camera mounted on the Unmanned Ground Vehicle, developed by Martin Marietta under a contract for the U.S. Government, captured a sequence of images as the vehicle moved along rough terrain in the countryside, thus undergoing continuously changing rigid motion, (a) shows one frame of the sequence with the normal flow field overlaid, (b), (d) and (f) show the positive (light color) and negative (dark color) vectors of the longitudinal patterns corresponding to the *x-, y-* and ;-axes. (c), (e) and (g) show the corresponding fitted patterns, (i) shows superimposed on the image the boundaries of the patterns whose intersections provide the FOE and the AOR. (j) Because measurements are not everywhere available (strong spatial gradients appear sparse) a set of patterns can possibly be fitted resulting in two bounded areas as solutions for the FOE and the AOR.

5 The Case of Fixating Stereo: Preliminaries

Consider an active binocular observer capable of fixating on an environmental point. The geometry of the system can be described as a constrained rigid motion, between the left and right eye. If we fix a coordinate on the left eye with the 2-axis aligned with its optical axis the y-axis perpendicular to the fixation plane, then the transformation relating the right eye to the left is a rotation around the y-axis and a translation in the *xz* plane (Figure 6). At the fixation point the disparity measurements are zero and in a neighborhood around it relatively small. Thus, it is legitimate to approximate the disparity measurements through a continuous velocity field. This amounts to the small baseline approximation that has been used in the literature [Carding *et a*/., 1993].

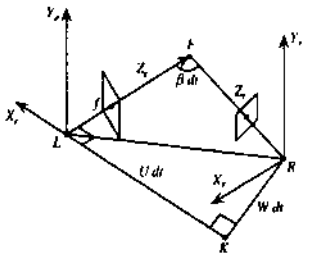


Figure 6: $LK = U\,dt$: translation along X_l-axis. $KR = W\,dt$: translation along Z_l-axis. $LFR = \beta\,dt$: angle denoting rotation around Y_l-axis. L, K, R, F belong to the fixation plane. dt is a hypothetical small time interval during which the motion bringing $X_lY_lZ_l$ to $X_rY_rZ_r$ takes place.

Denoting, as usual, by U and W the translation along the x- and z-axes and by β the rotation around the y-axis, and setting x_0 equal to $\frac{U}{W} \cdot f$ (x-coordinate of the focus of expansion, where f is the focal length), the component u_n of the disparity vector (u, v) along the gradient direction (n_x, n_y) is:

$$u_n = \frac{W}{Z}(-x_0 n_x + x n_x + y n_y) - \beta\left(f n_x + \frac{x^2}{f} n_x + \frac{xy}{f} n_y\right) \quad (1)$$

The exact geometry of the stereo configuration cannot be assumed and we do not wish to attempt the usual two step approach of first computing it from the available information in order to utilize it and derive in a second step the depth estimates. The reason is that small errors in the parameter estimation of the extrinsic geometry can result in large errors in the depth or shape estimates.

Our approach consists of two steps: (a) First, from one fixation and appropriately utilizing equation (1), we derive a set of partial depth orderings, classes of points for which we know the ordinal depth within each class. (b) From new fixations we obtain new partial orderings, though in different coordinate systems. The fixation geometry allows us to compare certain values between different fixations. This way we merge values in different fixations and build a global ordinal shape representation.

5.1 Ordinal Depth from One Fixation

An active binocular stereo system capable of changing its geometric parameters in a controlled vay should be aware of the pose of its eyes with regard to some head frame centered coordinate system. Thus it should know the angle the optical axis endoses with the baseline, which amounts to knowing the parameter x_0. If for a particular system this knowledge is not available, utilizing the constraints described in Section 4, the direction of the translation X_0 can be derived from the patterns of the normal disparity field, utilizing only the sign of the disparity measurements.

We do not know, however, the amount of rotation /3 and we also don't have to know the distance between the two eyes. Using equation (1) it is possible to obtain an ordinal depth representation for the scene whose image points lie on families of curves: Dividing equation (1) by $-n_x$ we obtain

$$-\frac{u_n}{n_x} = \frac{W}{Z}\left(x_0 - x - y\frac{n_y}{n_x}\right) + \beta\left(f + \frac{x^2}{f} + \frac{xy}{f}\frac{n_y}{n_x}\right) \quad (2)$$

We now classify normal disparity measurements according to their direction as a function of the image coordinates and a constant C. For this reason, we introduce a function $g(x, y, x_0)$ such that $x_0 - x - y\frac{n_y}{n_x} = g(x, y, x_0)$ and $(f + \frac{x^2}{f} + \frac{xy}{f}\frac{n_y}{n_x}) = C_i g(x, y, x_0)$ for some scalar C_i. This defines at each image point (x, y) a direction for the gradient (n_x, n_y)

$$\frac{n_y}{n_x} = \frac{C_i(x_0 - x) - f - \frac{x^2}{f}}{y\left(C_i + \frac{x}{f}\right)} \quad (3)$$

That is, if at a point (x, y) the gradient (n_x, n_y) is defined by equation (3) as a function of (x, y) and C_i, equation (2) can be written as:

$$\frac{-u_n}{n_x g(x, y, x_0)} = \frac{W}{Z} + \beta C_i \quad (4)$$

and thus we obtain a measurement linear in the inverse depth value. For any two points (x, y) and (x', y') in the image with depth Z and Z' and gradients (n_x, n_y), (n_x', n_y') defined by (3) with the same C_i, i.e., $\frac{n_y}{n_x} = \frac{C_i(x_0 - x) - f - \frac{x^2}{f}}{y\left(C_i + \frac{x}{f}\right)}$ and $\frac{n_y'}{n_x'} = \frac{C_i(x_0 - x') - f - \frac{x'^2}{f}}{y'\left(C_i + \frac{x'}{f}\right)}$ the depth order of the corresponding points can be derived. If $\frac{-u_n}{n_x g(x, y, x_0)} \gtreqless \frac{-u_n'}{n_x' g(x', y', x_0)}$ it follows that $Z \lesseqgtr Z'$.

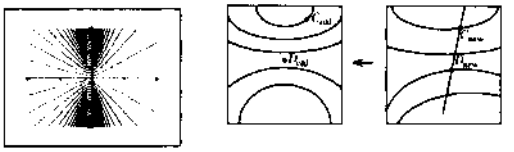
This result can be given the following geometric interpretation: equation (3) for every C_i defines the normals for a family of curves on the image. Each family of curves is defined by the following differential equation: $\frac{dy}{dx} = \frac{y\left(C_i + \frac{x}{f}\right)}{f + \frac{x^2}{f} - C_i(x_0 - x)}$. Its solution is:

$$y = \sqrt{\frac{x^2}{f} + Cx - Cx_0 + f} \cdot e^{\frac{C_i}{2}\lambda} \cdot \kappa$$

where

$$\lambda = \begin{cases} \frac{2}{\sqrt{q}} \cdot \tan^{-1}\left(\frac{\frac{2x}{f}+C}{\sqrt{q}}\right) & \text{if } q > 0 \\ \frac{-2}{\sqrt{-q}} \cdot \tanh^{-1}\left(\frac{\frac{2x}{f}+C}{\sqrt{-q}}\right) & \text{if } q < 0 \quad \& \quad \left|\frac{\frac{2x}{f}+C}{\sqrt{-q}}\right| < \\ \frac{1}{\sqrt{-q}} \ln \frac{\frac{2x}{f}+C-\sqrt{-q}}{\frac{2x}{f}+C+\sqrt{-q}} & \text{if } q < 0 \quad \& \quad \left|\frac{\frac{2x}{f}+C}{\sqrt{-q}}\right| > \end{cases} \tag{5}$$

with $q = 4 - \frac{4Cx_0}{f} - C^2$ and κ an arbitrary constant. Figure 7a shows such a family for $C_i = 0.2$, $f = 1$, $x_0 = 1$. Different C_i's provide different classes of curves. All points with the gradient perpendicular to the curves of one family can be ordered according to their depth.



(a) (b)

Figure 7: (a) Shape computation from one pair of images taken by a binocular fixating vision system: A partial ordering of the depth values can be obtained for all points with edges tangential (or normal disparity measurements perpendicular) to the curves of a family. (b) Ordinal depth information between C_{new} and D_{new} is used to obtain ordinal depth information between C_{old} and D_{old}.

5.2 Merging of Results from Different Fixations

A new fixation is obtained by rotating the left and the right eyes around the centers of their coordinate systems (their nodal points). Under a rotation the coordinates of every point $\vec{X}_{old} = (X_{old}, Y_{old}, Z_{old})$ in the original coordinate frame are related to the coordinates $\vec{X}_{new} = (X_{new}, Y_{new}, Z_{new})$ in the new coordinate frame after fixation through the linear transformation $\vec{X}_{new} = R \cdot \vec{X}_{old}$ (or $\vec{X}_{old} = R^T \cdot \vec{X}_{new}$), where R is a 3×3 rotation matrix, with coefficients $r_{11}, r_{12}, \ldots, r_{33}$ (and R^T the transpose of this matrix). The Z-coordinates Z_{new} and Z_{old} in the two different frames are thus related as follows:

$$Z_{old} = Z_{new}(r_{13}x_{new} + r_{23}y_{new} + r_{33}f) \cdot f \tag{6}$$

Points in the original image, whose coordinates are defined by a line equation

$$r_{31}x_{old} + r_{32}y_{old} + r_{33}f = K, \tag{7}$$

where K is a constant, are mapped to lines defined by the equation

$$r_{13}x_{new} + r_{23}y_{new} + r_{33}f = \frac{f}{K} \tag{8}$$

in the new image coordinate system.

For any two points (A, B) on these lines the ratio of their depth values is preserved under rotation. That is, if $Z_{A_{old}}$ and $Z_{B_{old}}$ are the depth values associated with the corresponding image point A_{old}, B_{old} in the original coordinate system and $Z_{A_{new}}$ and $Z_{B_{new}}$ are the depth values of the corresponding image point A_{new}, B_{new} then

$$\frac{Z_{A_{old}}}{Z_{B_{old}}} = \frac{Z_{A_{new}}}{Z_{B_{new}}} \tag{9}$$

This relationship allows us to merge ordinal depth maps from different fixations: In the new coordinate system (new fixation) we obtain a number of partial orderings for classes of points to be found on families of curves (as described in 5.1). Let us consider any such family of curves and let us intersect these curves with the lines defined by equation (8). For any new points to be found on the intersection of a line and one family of curves the ordinal depth relation obtained in the new coordinate system can be utilized to increase the knowledge about the ordinal depth maps in the original frame (see Figure 7b). Let C_{new} and D_{new} be two points with gradients perpendicular to the curves of one family and also lying on one line. These points correspond to C_{old} and D_{old} in the old coordinate system. C_{old} belongs to one family of curves and D_{old} to (usually) another family of curves in the old coordinate system Thus the depth measurements of C_{old} and D_{old} could not be compared using only one fixation. Since however, we know the ordinal depth in the new coordinate system and we also know the ratio of differences between a pair of points in the two coordinate systems, we can increase the number of points with ordinal depth information of the family of curves on which C_{old} lies, by the point D_{old}. Similarly, we can increase the number of points of the family containing D_{old}, by the point C_{old} [Fermüller and Aloimonos, 1995]. Any two measurements in the new coordinate system which lie on a straight line (given by equation (7)) and on the curves of one family can be utilized to increase the number of points in the partial orderings. Comparing all possible lines and performing a series of fixations, the partial depth maps in the first coordinate system successively can be built up.

It is obvious that during the process of successive fixation one could theoretically solve for β. If, in the frame at some fixation, two points lying on curves of one family and also lying on one of the straight lines are found which have equal depth, this information can be utilized to solve for β in the original frame. Such an approach, however, is in contradiction with the approach advocated in this paper. Shape computations are local computations, and thus should be done in such a way that does not allow error sources to propagate. If β were computed, a small error in it would contaminate the shape computations everywhere in the image.

Our computational strategy utilized information about X_0 (i.e., the angle formed between the x-axis of the left eye and the baseline). Although the theory described in Section 4 can be used to derive this information, the system could derive x_0 from motor information without involving the imagery; however, *(3* can only be computed from image information. Its derivation requires exact fixation (i.e., the exact intersection of the two cameras' optical axes on a surface point), which is practically impossible. Figure 8 describes experimental results with an active binocular head/eye system.

Finally, it should be noted that instead of computing an ordinal representation in the depth *Z,* we could derive an ordinal representation in the distance *R* from the nodal point [Fermuller and Aloimonos, 1995]. If we use spherical eyes and employ a spherical coordinate system (r, θ, ϕ) we obtain similarly as in section 5.1 families of curves defined as functions in θ, ϕ, x_0, and C, along which ordinal distance information can be derived. Such a representation provides an advantage in the phase of merging data from different fixations. Since the distance *R* remains the same in the different coordinate systems of different fixations (which are only rotated to each other), any two distance measurements on the same family of curves in any fixation and not only those on lines can be used to obtain additional distance information in the original frame.

6 Conclusions

In the past few years, it has become clear that Vision (and perception in general) should not be studied in isolation but in conjunction with the physiology and the tasks that systems perform. In this paper we argued that the synthesis of vision and action must happen through spatiotemporal representations computed from well defined input. We showed how an active observer by appropriately selecting subsets of the input can estimate representations of motion and structure through pattern matching.

7 Acknowledgements

The support of the National Science Foundation under Grant IRI-90-57934 and the Austrian "Fonds zur Forderung der wissenschaftlichen Forschung" Project No.S7003 is gratefully acknowledged.

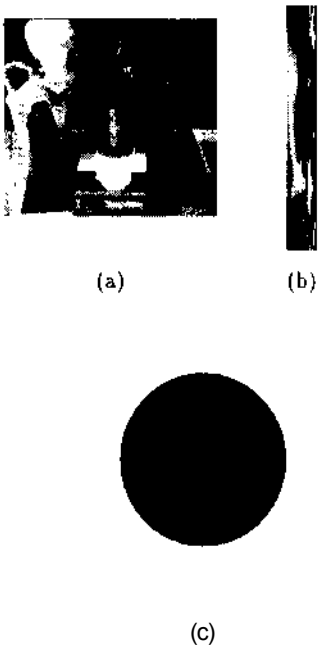


Figure 8: Computation of ordinal depth map: A binocular stereo system performed fifteen successive fixations on various object points in the scene shown. The images were transformed to logpolar representations, which changes the resolution within the image und thus allows to derive normal disparity measurements with differential techniques within a bounded area around the image center, (a) first image taken by the left camera and the circular area for which a depth map was computed, (b) log polar transform of the first image, (c) ordinal depth map computed after fifteen fixations. The depth values computed at edges are coded as grey values, where white denotes the nearest and black the furthest point. Areas in which no edges were found are colored in a uniform grey.

References

[Aloimonos *et al,* 1988] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal of Computer Vision,* 2:333 356, 1988.

[Bajcsy, 1988] R. Bajcsy. Active perception. *Proceedings of the IEEE,* 76:996-1005, 1988.

[Ballard and Brown, 1992] D.H. Ballard and CM. Brown. Principles of animate vision. *CVGIP: Image Understanding: Special Issue on Purposive, Qualitative, Active Vision, Y. Aloimonos (Ed.),* 56:3-21, 1992.

[Faugeras, 1992] O. Faug eras. *Three Dimensional Computer Vision.* MIT Press, Cambridge, MA, 1992.

[Fermuller and Aloimonos, 1995] C. Fermuller and Y. Aloimonos. Estimating ordinal shape representations. Technical Report CAR-TR, Center for Automation Research, University of Maryland, 1995.

[Garding do/., 1993] J. Carding, J. Porrill, J.E.W. Mayhew, and J. P. Frisby. Binocular stereopsis, vertical disparity and relief transformations. Technical Report TRITA-NA-P9334, CVAP, Royal Institute of Technology, Stockholm, Sweden, 1993.

[Koenderink and van Doom, 1991] J.J. Koenderink and A.J. van Doom. Affine structure from motion. *Journal of the Optical Society of America,* 8:377-385, 1991.

[Todd and Reichel, 1989] J. Todd and Reichel. Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychology Review,* 96643—657, 1989.