COMPUTATION OF MOTION IN DEPTH PARAMETERS: A FIRST STEP IN STEREOSCOPIC MOTION INTERPRETATION

P. Balasubramanyam
M.A. Snyder

COINS Technical Report 88-50

May 1988

COMPUTATION OF MOTION IN DEPTH PARAMETERS: A FIRST STEP IN STEREOSCOPIC MOTION INTERPRETATION.

Poornima Balasubramanyam

M.A. Snyder

April 1988

ABSTRACT

Abstract

This paper[1] is motivated by the need to obtain the three parameters of absolute 3D motion in depth of objects in a dynamic scene using visual information. *Motion in depth*

[1]Submitted to the Darpa IU Workshop to be held April 6 - 8 at Cambridge, MA

(MID) parameters refer to the following three components of the object or camera motion, i.e., the single component of translation in depth (parallel to the line of sight) and the two components of rotation in depth (or rotational components that are not about the line of sight). In this paper we show that use of *stereoscopic motion* enables the MID parameters to be computed in a quick and robust manner for a stereo camera system moving through an environment that may have several independently moving rigid bodies. It has been shown elsewhere that use of temporal binocular imagery permits the formulation of the ratio of the relative optic flow between a stereo pair of images and the disparity between them as a *linear function* of the image coordinates. The coefficients of this linear function are the translation in depth and rotation in depth components which are precisely the MID parameters being computed in our approach.

This work has been supported by the following grants: DARPA under grant N00014-82-K-0464 and DARPA and RADC under contract F30602-87-C-0140.

1. Introduction

Motivation

The human visual system exhibits remarkable skill in detecting the *translation in depth* of moving objects. Consider, for instance, the following two commonly performed tasks – catching a ball that is thrown toward the observer with a rotational spin component contributing to the motion, and navigating around obstacles moving toward or away from the observer. The observer has to make rapid judgements about the translation in depth of the object in question. Determination of the complete trajectory of the object involves determination of all the parameters of motion and is less important to the immediate needs of the observer. It is possible to make such judgements of translation in depth even when the object has other motion attributes, such as rotations about different axes contributing to the actual trajectory. Perception of the translation in depth of the spinning ball moving toward the observer is one such example. Hence, we believe that the translation in depth of objects is a particularly useful motion parameter to compute for both navigation and moving obstacle avoidance.

Also of significant interest to us is the psychophysical evidence that demonstrates the perception of *rotation in depth* by human observers. There are specific empirical findings [10,18,20] that are concerned with the ability of human observers to perceive three dimensional relationships on the basis of rotation in depth. The emphasis in the above work has been on both the perception of the direction of rotation as well as on the overall impression of depth or coherence elicited by the rotating objects. Computational models dealing with the structural information that can be extracted from such an image transformation have been developed [29,30]. We believe that the perception of rotation in depth of objects in the image can be used to guide further semantic analysis of the scene in order to yield structural as well as motion descriptions of semantically meaningful objects in the scene.

Motivated by this, we propose here a model of *motion in depth* computation that permits the presence of general motion of both the camera and multiple objects in the environment.

Approach

Motion in depth (MID) parameters refer to the following three components of the object or camera motion: the single component of translation in depth (along the line of sight) and the two components of rotation in depth (rotations that are not about the line of sight).

The model uses stereo time-varying optic flow fields to perform a fast and robust computation of the parameters that specify the absolute MID values. It is based on the concept of stereoscopic motion, described by Beverly and Regan [9]. *Stereoscopic motion* refers to the relative motion present between the stereo pair of a temporal image sequence. Psychophysical evidence [9] indicates that stereoscopic motion is used by human observers to extract the translation in depth of objects in the scene.

In the proposed model, therefore, we consider the *dynamics* of a scene as viewed by a stereo pair of cameras. This is distinct from the analysis of the static imagery of the stereo pair at one time instant to obtain the static disparity between the two images. The model assumes for its input, the existence of the static disparity field corresponding to the first stereo pair of the two frame stereo sequence along with the optic flow computed seperately for each of the two stereo image sequences. These are used to determine the relative image change due to motion in the temporal binocular image sequence. Issues dealing with the correspondence problem in stereo and motion analysis are bypassed for the purposes of the current work. Hence, the model proposes an integrated *interpretation* of the disparity and optic flow information as opposed to an integrated analysis of stereo and motion that deals with correspondence issues. The latter approach has been adopted in other work in this area [14,31]. These and other related work have been discussed in the following section.

It was shown by Waxman and Duncan [31] that use of temporal binocular imagery permits the formulation of a linear relationship between the relative flow and disparity values and the MID parameters. This was used to establish correspondence between stereo pairs of images. We believe this to be insufficient use of the information that is available from the linear relationship and hence propose to use this formulation as the theoretical basis for our approach to computing the MID parameters of object or sensor motion for a sensor travelling through an environment that may have independently moving objects.

A two-stage global technique is employed here to compute the three motion in depth parameters. In the first stage, we employ the segmentation technique developed by Adiv [1,2] as the initial stage of his algorithm for interpreting monocular optic flow fields. We perform this segmentation in order to obtain a 2-D grouping of the optic flow field. In the second stage, segments from the first stage are grouped into regions that correspond to the same set of three MID parameters by employing a least squares minimization technique on the relative optic flow between the stereo pair of image sequence.

Note that such a two-step view of 3-D motion interpretation essentially follows the viewpoint of Adiv [1]. Performing segmentation on the 2D image plane provides a method of restricting the actual 3D interpretation mechanism, (i.e., the global minimization step in the current model), to a semantically relevant set of flow vectors and thus contributes to the robustness of the entire algorithm. A direct attempt at interpreting the 3D motion from the 2D flow without an intermediate grouping would necessarily be a local analysis of the flow and hence be very susceptible to noise. On the other hand, it is important to note that this segmentation step uses an approximation to determine the grouping. Hence, it is important not to use this step to directly compute the values of the 3D motion parameters, but use this grouping as a mask on the *relative flow field* in order to determine the values of the MID parameters without making any surface approximations.

MID parameters can be obtained from a general 3D motion computation algorithm using monocular imagery, but these algorithms are computationally intensive and do not permit quick reliable computation. One of the chief reasons for this is that the flow information to the depth and motion parameters via nonlinear equations. The main advantage provided by the proposed integrated interpretation of stereo and motion information for extracting motion in depth is that stereo information provides additional constraints that make it possible to formulate a *linear* relationship between the data in the flow and disparity fields and the motion in depth parameters. This, in turn, makes it possible to devise a direct computation without hypothesizing any of the motion parameters such as in [1,21]. This can conceivably be used to directly compute the remaining three motion parameters, again, without employing a hypothesize and test scheme. Hence, extracting all the motion parameters would become a problem of handling two sets of linear functionals, and we plan to demonstrate this in future.

In Section 2, we review existing techniques for dealing with the problem of integrating stereo and motion information. The assumptions and limitations of these techniques are discussed. We also briefly review current motion interpretation research. In Section 3, we formulate the theoretical model. In Section 4, we describe the algorithm and discuss the decisions that were made in the development and implementation of the algorithm. Experiments based on simulated data are described in Section 6. These experiments demonstrate the generality of the algorithm. In future work, we plan to demonstrate the algorithm on real data. Anandan's algorithm [4,5] which has shown state of the art performance on real imagery, will be used to extract the optic flow fields. A modified version of the same algorithm can be used to extract the disparity values as well. In Section 7, we summarize the approach and major results, and discuss directions for future research.

2. Literature Survey

In this section we examine the approaches taken by several authors to the problem of interpreting temporal binocular imagery. The research over the years has chiefly been on the seperate interpretation of optic flow in monocular imagery or on static stereo analysis. An extensive review of work on the interpretation of monocular optic flow is given in [8], and more recent work includes that of [3,32].

Any method for the interpretation of monocular optic flow fields is limited by the fact that the flow value at a point in the image is a nonlinear function of the six parameters of motion and the depth of the corresponding environmental point. Dealing with this imposes restrictions on any interpretation technique. For instance, iterative methods such as [11,21,27] need good initial guesses. Sensitivity to noise is reported by much of the earlier research e.g.,[11,21,26,28]. Some techniques, such as [16] and [12], either assume restricted motions such as pure translation or assume that one component of the motion such as the rotation may be known and can be subtracted out prior to the computation. Others deal with a stationary environment and moving camera, disallowing the possibilty of multiple independently moving objects [25].

Some of the more successful general methods that deal with multiple independently moving objects, as well as the presence of general motion, can be found in [1,2] and [32]. Good results in determining 3-D motion and object structure for simulated data have been reported in [32], although computation of the local derivatives of the flow may be highly sensitive to noise.

The work of **Adiv** [1,2] appears to be robust. However, the technique adopted for the computation of the motion parameters is computationally intensive and does not permit quick evaluation of at least some of the motion parameters that would be important in a practical context. This is a defect that is inherent to any method that deals with monocular imagery due to the underconstrained nature of the relationship between the flow field and the information desired from it. In the case of the method adopted in [1,2], for instance, the motion parameters are computed only at the end of a search of a sampling of possible translation directions and corresponding optimal rotation parameters, an approach that is essentially a bottom-up hypothesize and test scheme in practice.

Any method that is directed toward quickly extracting motion parameters, while retaining the ability to deal with multiple objects and general motion has to deal with the fact that monocular imagery with just two frames provides underconstrained information. There is a need to look for additional sources of information that will provide more constraints for computing the motion parameters. This leads to the use of

- multiple frames of monocular imagery, and
- stereo time-varying imagery.

We now briefly review the previous work that addresses the latter aspect of integrating stereo and motion analysis. Techniques have been formulated to facilitate the solution to the correspondence problem for both stereo and motion as well as to address interpretation issues.

The first use of the term *stereoscopic motion* was in the work of **Regan, et al.** [22]. They addressed the problem of using visual information to recover motion in depth of objects, and provide evidence to support models of neural organizations in the human visual system for detecting the motion in depth of objects. They propose the presence of neural "filters" that are sensitive to the relative velocities in the left and right retinal images and are thereby selectively sensitive to the *direction* of motion in depth of objects in the visual field. These motion-in-depth detectors are thus viewed as binocularly driven channels that process the changing disparity. Our computational model has been motivated by the psychophysical evidence that strongly supports such a formulation. Also of interest to us is pychophysical evidence in [22] for the following –

- changing disparity grows relatively more effective as the velocity increases, and
- changing disparity grows relatively more effective as the inspection time increases.

In future work, we will examine the relevance of these conclusions to the proposed model.

Other work of Regan [23] indicates that the human visual system possesses sensitivity to four kinds of relative motion, namely,

- the velocity difference between two points in one retinal image along the line joining the two points,
- the velocity difference between the two points in one retinal image perpendicular to the line joining them,
- rotary motion, and
- the ratios of the velocities of the left and right retinal images of an object moving in depth.

We note that the latter sensitivity may be used to recover object motion in depth. An interesting alternate viewpoint provided here is that these filters may provide physiological means of analysing the local flow in the retinal images in order to recover specific information about the environment.

The approach taken by **Richards** [24] to integrate stereo and motion information uses both to extract three dimensional information about the environment in a manner that results in a unique solution for the *object structure*. The problem with solving for structure from pure motion information is that any algorithm needs to deal with second order equations relating the flow to the structure. This will result in the possibility of multiple interpretations of object configurations. It is required that these be removed by disambiguating the possible solutions. The problem noted with stereopsis is that the correct configuration of objects can be determined only if the fixation distances are known. This is because the same configuration at different distances will result in different angular disparities. Motion information can correctly interpret the angular relations between objects, thus making knowledge of the fixation distance unnecessary. The approach uses stereopsis to provide information which disambiguates the multiple interpretations found with pure motion, since stereo can provide absolute depth information. We note that this work deals purely with the problem of extracting the structure of the environment and does not deal with the use of stereo and motion information to extract the motion parameters of objects in the environment or of the camera.

The approach of **Jenkin** [14] uses stereo and motion information in a prediction–correction mechanism to facilitate the solution to the *correspondence problem* both in stereo and motion analysis. The algorithm predicts the position that a point in the current frame might correspond to, using the previous motion of that point, i.e., by using the previous frames. Hence, the stereopsis correspondence problem is simplified since the search area is restricted to that predicted by the analyses of the previous frame pairs. We note that the algorithm addresses only the correspondence problem for both stereo and motion, and not the interpretation of stereo or motion information.

A unified approach to the analysis of stereo and motion data is given by **Waxman and Duncan** [31]. They show that there is a correlation between the stereo disparity and the relative flow between the stereo pair of an image sequence. This result is used to establish correspondence in the context of local support from neighbourhoods. It is proposed that in the analysis of time–varying stereo imagery, after the initial correspondence is established, matching for subsequent images need be performed only at the peripheral regions of the image as well as around occluding boundaries. We note that this work approaches the integration of stereo and motion information from the viewpoint of facilitating the correspondence problem.

An entirely different approach using two frames of stereo imagery is taken by **Huang and Blostein**, [13]. They estimate the rigid body rotation and translation parameters by matching 3-D points determined at two time instances from stereo information. This 3–D matching problem can be solved by considering geometric relationships that should be preserved under rigid body displacements. The 2–D matching problem continues to persist for the stereo matching, while the motion estimation algorithm needs to deal with appropriate mechanisms for 3–D matching.

Mutch [19] describes a technique to recover the translation in depth of a moving object point, using the concept of stereoscopic motion as described by [9]. The change in the location of the image of an object point is defined by its "change vector." The relative difference between two change vectors in the left and right image sequences corresponding to a translating object point is such that their relative orientation and magnitude leads to the perception of certain 3–D properties of the object, including its translation in depth. The direction of the translation in depth as well as the position of the impact point can be determined from this method. We note the approach requires that in the case of general motion by the object point, the rotational component first be removed from the change vector. This work is of interest since it uses stereoscopic motion in a computational context to determine the translation in depth of objects.

We give a more general approach to the problem of interpreting stereo and motion information by being able to deal with general motion and the presence of multiple independently moving objects. Also, we do not restrict the point of interest to a single object point, such as the centroid in [19], but develop the technique for dense flow and disparity fields. This is more robust since more global information is allowed in the computation. Finally, we extract translation as well as rotations in depth of the camera and the objects.

3. Theoretical Formulation Of The Model

In this section we develop the mathematical framework for our approach. We first consider monocular flow analysis, and then generalize binocular flow.

Monocular Flow Analysis

Given the optic flow on the image plane, we can relate the values of the components of the flow field at every point in the image to the 3–D motion parameters and depth of the environmental point that projects to this image point.

Let us consider a cartesian coordinate system (X, Y, Z) that is fixed with respect to the camera with the focal length normalized to a distance of 1 from the origin, O, to the image plane. Let (x, y) represent the image coordinate system. The perspective projection of an environmental point, (X, Y, Z), on the image plane is then given by

$$x = X/Z, \tag{1}$$

$$y = Y/Z. \tag{2}$$

We now consider the motion relative to the camera of a rigid object in the environment. Let P be the position vector of some point on the object, with camera coordinates given by (X, Y, Z) (see Figure 1). Since the object is rigid, the instantaneous velocity, $\dot{P}$, of P can be represented as an combination of an infinitesimal rotation Ω, and an infinitesimal translation T. Thus,

$$\dot{P} = \Omega \times P + T. \tag{3}$$

In components, this becomes :

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = \begin{pmatrix} \Omega_Y Z - \Omega_Z Y + T_X \\ \Omega_Z X - \Omega_X Z + T_Y \\ \Omega_X Y - \Omega_Y X + T_Z \end{pmatrix}.$$

The corresponding image point (x, y) has an image velocity given by (α, β), where

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \frac{1}{Z^2} \begin{pmatrix} \dot{X}Z - \dot{Z}X \\ \dot{Y}Z - \dot{Z}Y \end{pmatrix},$$

which becomes, from (3) :

$$\alpha = -\Omega_X xy + \Omega_Y(1 + x^2) - \Omega_Z y + (T_X - T_Z x)/Z, \tag{4}$$

$$\beta = -\Omega_X(1 + y^2) + \Omega_Y xy + \Omega_Z x + (T_Y - T_Z y)/Z. \tag{5}$$

We therefore see that the flow equations (4,5) are 2^{nd} order functions of (x, y).

Binocular Flow Analysis

We now extend the above derivation to the case of a stereo pair of cameras :

The two image planes are coplanar,and perpendicular to the ground plane. We assume that the focal points lie along the same horizontal line (see Figure 2). The analysis follows that of [31].

Let us denote the cartesian coordinate system for the left camera by (X_l, Y_l, Z_l) and that for the right camera by (X_r, Y_r, Z_r). Let the horizontal displacement between the two focal points, O_l and O_r (see Figure 2) be b. We denote entities with respect to the left and right cameras by the subscripts l and r respectively. The final formulations are with respect to the left coordinate system.

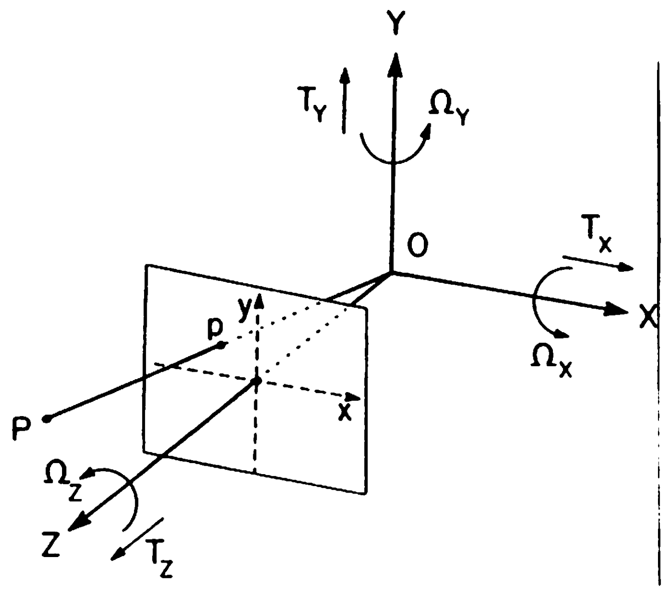

Figure 1: Monocular Camera Configuration.

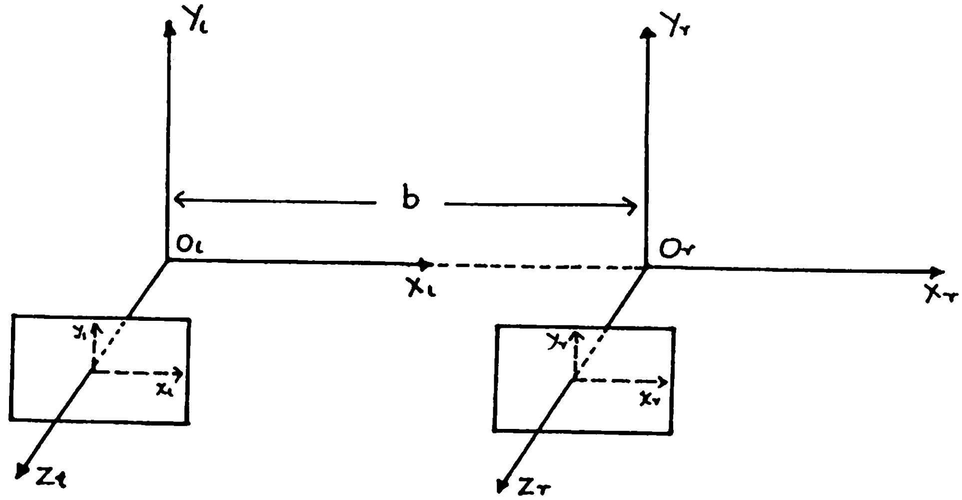

Figure 2: Binocular Camera Configuration

Consider a point at (X_l, Y_l, Z_l). The projection of this point on the left image plane is given by (x_l, y_l). Similarly, the point is at a position (X_r, Y_r, Z_r) with respect to the right coordinate system, and it projects on the right image plane at (x_r, y_r). Let the disparity between the corresponding image points be given by δ, where

$$\delta \stackrel{\text{def}}{=} x_l - x_r = b/Z, \tag{6}$$

since

$$Z_l = Z_r \equiv Z. \tag{7}$$

Also note that

$$y_l = y_r \equiv y. \tag{8}$$

Let us now consider the flow equations for the two cameras. From equations (4,5), we have with respect to the left camera

$$\begin{aligned} \alpha_l(x_l, y_l) &= -\Omega_{X_l} x_l y_l + \Omega_{Y_l}(1 + x_l^2) - \Omega_{Z_l} y_l + \\ &\quad (T_{X_l} - T_{Z_l} x_l)/Z, \end{aligned} \tag{9}$$

$$\begin{aligned} \beta_l(x_l, y_l) &= -\Omega_{X_l}(1 + y_l^2) + \Omega_{Y_l} x_l y_l + \Omega_{Z_l} x_l + \\ &\quad (T_{Y_l} - T_{Z_l} y_l)/Z, \end{aligned} \tag{10}$$

and with respect to the right camera,

$$\begin{aligned} \alpha_r(x_r, y_r) &= -\Omega_{X_r} x_r y_r + \Omega_{Y_r}(1 + x_r^2) - \Omega_{Z_r} y_r + \\ &\quad (T_{X_r} - T_{Z_r} x_r)/Z, \end{aligned} \tag{11}$$

$$\begin{aligned} \beta_r(x_r, y_r) &= -\Omega_{X_r}(1 + y_r^2) + \Omega_{Y_r} x_r y_r + \Omega_{Z_r} x_r + \\ &\quad (T_{Y_r} - T_{Z_r} y_r)/Z. \end{aligned} \tag{12}$$

But from the geometry we know that

$$\Omega_l = \Omega_r \equiv \Omega \tag{13}$$

and

$$T_r = T_l - \Omega_l \times b\hat{x}. \tag{14}$$

where $\hat{x}$ is the unit vector pointing from O_l to O_r. Hence we can rewrite the flow equations (11,12) for the right image plane using equations (13, 14). We then have with respect to the left coordinate system,

$$\begin{aligned}\alpha_r(x_l+\delta,y_l) &= -\Omega_{X_l}(x_l y_l+\delta y_l)+\Omega_{Y_l}(1+\delta x_l+x_l^2)- \\ &\quad \Omega_{Z_l}y_l+(T_{X_l}-T_{Z_l}\delta-T_{Z_l}x_l)/Z \quad (15)\\ \beta_r(x_l+\delta,y_l) &= -\Omega_{X_l}(1+y_l^2)+\Omega_{Y_l}x_l y_l+\Omega_{Z_l}x_l+ \\ &\quad (T_{Y_l}-T_{Z_l}y_l)/Z. \quad (16)\end{aligned}$$

The two components of the ***relative flow*** between the two images are thus given by

$$\begin{aligned}\Delta\alpha &= \alpha_r(x_l+\delta,y_l)-\alpha_l(x_l,y_l)= \\ &\quad [\Omega_{Y_l}x_l-\Omega_{X_l}y_l-T_{Z_l}/Z]\delta, \quad (17)\\ \Delta\beta &= \beta_r(x_l+\delta,y_l)-\beta_l(x_l,y_l)=0. \quad (18)\end{aligned}$$

We can now write the ratio of the relative flow between the two image points to the disparity, expressed with respect to the left camera coordinate system, as :

$$\frac{\Delta\alpha}{\delta} = \Omega_Y x_l-\Omega_X y_l-T_Z/Z \quad (19)$$

$$\frac{\Delta\beta}{\delta} = 0. \quad (20)$$

We refer to the vector $(\frac{\Delta\alpha}{\delta},\frac{\Delta\beta}{\delta})$ as the ***relative flow vector***, and a field of these vectors as the ***relative flow field***. Since only the "left" quantities appear in the remainder of this work, we drop the "*l*" subscript to denote this.

An interesting point derived in [31] is that the x-component of the relative flow vector, given by equation (19), is identical to the ratio of the rate of change in the disparity to the disparity. We now derive this relationship.

We know that the disparity, δ, is given in equation (6) as

$$\delta=\frac{b}{Z} \quad (21)$$

Hence, the rate of change of disparity, $\dot{\delta}$, is given as

$$\dot{\delta}=-\frac{b}{Z^2}\dot{Z}=-\delta\frac{\dot{Z}}{Z} \quad (22)$$

Referring to equation (3), we have

$$\frac{\dot{Z}}{Z} = \frac{1}{Z}(-\Omega_Y X + \Omega_X Y + T_Z) = -\Omega_Y x + \Omega_X y + T_Z/Z \tag{23}$$

Substituting this into equation (22), we find

$$\dot{\delta} = [\Omega_Y x - \Omega_X y - T_Z/Z]\delta \tag{24}$$

or

$$\frac{\dot{\delta}}{\delta} = \Omega_Y x - \Omega_X y - T_Z/Z. \qquad Q.E.D \tag{25}$$

Hence, we can obtain the relative flow field data in one of two ways, i.e.,

- take the relative flow between the two optic flow fields corresponding to the left and right cameras, or
- approximate the differential of the disparity field over the interval between the two time instances.

We choose to obtain the relative flow fields using the former method. This view is supported in in a psychophysical context since experimental studies [23] have shown some subjects to have areas of the visual field that are *blind* to static disparity and yet possess normal sensitivity to motion in depth. Hence, we believe that the extraction of motion in depth by computing the relative optic flow between a stereo pair is a closer model of the human perception of such motion.

4. Computation of the MID parameters using Stereoscopic Motion Constraints.

A Brief Introduction to the Approach

We employ a global technique to extract the three MID parameters $\{\Omega_X, \Omega_Y, T_Z\}$ in the presence of general motion of the camera and several independently moving objects.

The theoretical basis for this approach to the integration of the stereo and motion information was given in Section 3. Briefly, if u is the relative optic flow at image points in the left and right cameras corresponding to a single world point and δ the disparity between the left and right images at any one time instant for the same point, it was shown that the ratio of u to δ is a linear function of the image coordinates. The coefficients of this linear functional are the rotations about the X and Y axes and the translation along the Z axis, which are precisely the MID parameters being computed in our approach.

Algorithm Description

The inputs to the algorithm are the two flow fields, one each for the left and the right images, and the disparity field between the left and the right images at any one time instant.

The algorithm proceeds in the following three steps :

- Extract the relative flow field between the left and right images using the difference of the two optic flow fields along with the disparity field.

- Employ Adiv's technique, [1,2], for obtaining the segmentation of the monocular optic flow field corresponding to the left camera. Thus, we perform a segmentation of the optic flow field from pure motion information on the 2-D image plane in order to obtain a grouping of the vectors, where each segment corresponds to the motion of a roughly planar surface. We discuss the effect of performing a similar grouping on the relative flow fields in Section 4.1.

- Merge the segments on the 2D image plane (obtained from the previous segmentation step) based on a least squares minimization to compute the MID parameters for each of the merged regions. The output at this stage is a grouping of the image into regions that correspond to the same set (within some normalized value of the deviation) of three MID parameters, i.e., the two parameters of rotation in depth and the scaled translation in depth.

In the interest of robustness, the grouping obtained in the segmentation step is used as a template to guide the areas in the image where the minimization is employed. The actual optimization constraint is applied to the information in the relative flow field within each of a set of single or possibly merged group of segments in order to extract the MID parameters for them.

Also, since we deal only with MID parameters, the merged groups of regions *cannot* be interpreted as representing objects in the scene. They represent those regions in the image that have the same set (within some normalized value of the deviation) set of MID parameters. See Expt.4 (Figure 5e) in Section 5 for an example of a grouping wherein the background plane and a stationary ellipsoid in the environment get merged as one region simple because both have the same MID parameters relative to the camera.

We describe these two steps of the algorithm in the following two subsections.

4.1 Segmentation

At this stage of the algorithm, the optic flow field for the left camera is used to obtain a 2–D grouping of those flow vectors that are consistent with the motion of an approximately planar patch [1,2]. We also discuss the effects of performing a similar segmentation on the *relative flow field*.

Formulation Of The Segmentation Constraint

Let us first examine the use of optic flow for segmentation. We briefly review the segmentation process as developed in [1,2]. The viewer is advised to see these for more details. If we consider the flow field that is induced by the motion of a rigid planar surface described by

$$k_x X + k_y Y + k_z Z = 1. \tag{26}$$

Then we can rewrite equations (4,5) as :

$$\alpha = a_1 + a_2 x + a_3 y + a_7 x^2 + a_8 xy, \tag{27}$$

$$\beta = a_4 + a_5 x + a_6 y + a_7 xy + a_8 y^2, \tag{28}$$

where $\{a_1, \ldots, a_8\}$ are functions of the 3-D motion parameters, (T, Ω), of the objects in the environment (or conversely, the camera) and the three planar surface parameters, (k_x, k_y, k_z) :

$$a_1 = \Omega_Y + k_z T_X,$$

$$a_2 = k_x T_X - k_z T_Z,$$

$$
\begin{aligned}
a_3 &= -\Omega_Z + k_y T_X, \\
a_4 &= -\Omega_X + k_z T_Y, \\
a_5 &= \Omega_Z + k_x T_Y, \\
a_6 &= k_y T_Y - k_z T_Z, \\
a_7 &= \Omega_Y - k_x T_Z, \\
a_8 &= -\Omega_X - k_y T_Z.
\end{aligned} \tag{29}
$$

The desired output is an organization of the optic flow field into *segments*, where each segment corresponds to the motion of a roughly planar surface. The constraint used to perform this grouping is that each segment thus formed be consistent with a single Ψ-transformation {equations (27, 28)}. The Ψ-transformation space corresponds to the coefficients of the second order flow equations as functions of image coordinates. This space is 8-dimensional, and searching it for the Ψ-transformation that is consistent with the motion in the flow field would be very expensive computationally. Hence, the consistency constraint is approximately implemented by a two-step process :

- As a preprocessing step, the flow vectors are first grouped based on consistency with the 6-parameter linear approximation to equations 27,28 (an affine transformation) using a modified Hough transform. This yields *components*.
- These components are then merged together based on the minimization of an error measure derived using the full optimal Ψ-transformation.

We note several interesting features of this stage. The six parameter affine transformation corresponds to a Ψ-transformation with $a_7 = a_8 = 0$. In addition, the flow components $\alpha(\beta)$ depends only on the parameters $\{a_1, a_2, a_3\}(\{a_4, a_5, a_6\})$, i.e., the two components are decoupled. Because of this, the significant computational cost of a Hough transform on the 6-dimensional affine space can be mitigated by instead performing a seperate Hough transform on each of the two three-parameter affine spaces $\{a_1, a_2, a_3\}$ and $\{a_4, a_5, a_6\}$. In addition the transform is implemented on the 3-parameter affine spaces in a multi-resolution scheme that makes use of the concept of dynamically quantized spaces wherein the transform is iteratively employed around the value estimated in the previous iteration using a finer resolution.

It is important to note that this segmentation step uses an approximation to determine the grouping. Hence, the values of the Ψ-transformation themselves are not used to compute the motion parameters. This grouping is used as a mask on the *relative flow field* in order to determine the values of the MID parameters without making any surface approximations, in order to increase the robustness of the algorithm.

Difficulties with direct use of the relative flow field for segmentation

We now discuss our reasons for using the pure motion flow fields, rather than the relative flow field, for segmentation purposes. Using the planar patch approximation (26) in the relative flow equations (19,20), we find

$$\frac{\Delta\alpha}{\delta} = b_0 + b_1 x + b_2 y, \tag{30}$$

$$\frac{\Delta\beta}{\delta} = 0, \tag{31}$$

where

$$\begin{aligned} b_0 &= -T_Z k_z, \\ b_1 &= \Omega_Y - k_x T_Z, \\ b_2 &= -\Omega_X - k_y T_Z. \end{aligned} \tag{32}$$

Upon comparing equations (29) and (32) we see that the second order components of the *optic* flow field, i.e., the coefficients a_7 and a_8, are precisely the first order components of the *relative* flow field, namely, b_1 and b_2 respectively:

$$\begin{aligned} a_7 &= b_1, \\ a_8 &= b_2 \end{aligned} \tag{33}$$

Grouping using the relative flow constraint would thus result in a set of relative flow vectors that are consistent with the same set of three parameters, b_0, b_1 and b_2 which amounts to dealing only with the second order components of the optic flow field, and disregarding the first order components. This creates incorrect grouping since in situations where the motion is such that the second order components of the flow field are very small, the relative flow field is going to be a function of very small first order coefficients. This occurs in situations where the translation and rotations in depth are small. A Hough transform on such a space will create false peaks in the parameter space and thus be unreliable. Thus for purposes of grouping, we would like to utilise all the available information and obtain as reliable a grouping as possible.

Hence, we use the optic flow (corresponding to the left camera) in order to obtain the segmentation mask to be used in the optimization on the relative flow field. In the current implementation the optic flow corresponding to the left camera is chosen to obtain the segmentation.

4.2 Optimization

Segments that are consistent with the same set of three MID parameters are merged together in this stage, which proceeds in several steps:

- As an initial step, optimal MID parameters and a related error measure (see 35) are computed for each of the segments obtained from the segmentation step. The MID parameters are computed using a least squares error minimization (see 34).
- Sets of segments are sequentially tested for merger by deciding if the relative flow field in them corresponds to a single set of MID parameters.
- The merging decisions are based on the degree of consistency of the relative flow vectors in the entire set of segments being tested for merging to the same set of three optimal MID parameters. This is done by comparing the error measure obtained by taking the entire set of segments with the error measures for each of the segments taken singly. Both the error measures are computed with respect to the three optimal MID parameters that are obtained for the merged set using a least squares minimization.
- Only segments that have not been included in any previous merging are included in the next merging
- The MID parameters are computed for the merged sets of segments.
- All segments that correspond to the stationary environment are grouped together as one merged segment.
- Independently moving objects are picked out, unless the flow corresponding to the regions in and around the objects is such as to produce ambiguities during the interpretation process.
- The computation at this stage is *linear* in complexity since the technique considers only the *neighbouring* segments for merging decisions. This is reasonable since we are searching for groups in the image that correspond to the same set of MID parameters rather than recognise object masks.

Minimization Process

It is required that the MID parameters be extracted in the simplest possible manner that preserves robustness in the entire computation. In general, it is possible to obtain them for each segment from the values of the relative flow field at three points in that segment. We could then think of ways to merge neighbouring segments that exhibit the same MID parameters. But two factors need to be considered First, the flow fields and the disparity field are generally prone to corruption by noise. Second, the current method of obtaining the relative flow field by taking the difference of two flow fields adds numerical errors to the data. Given these two sources of data distortion, it is important to use all possible information in computing the parameters. A least squares error formulation is well suited for this purpose since it is more global in nature and robust in implementation.

Computation of the MID parameters

The least squares formulation minimizes the error between the actual relative flow field value at a point and that predicted by the MID parameters and the depth of the point. Hence, given a set of flow vectors, it selects the optimal set of three MID parameters that are consistent with the minimal deviation in the relative flow field predicted by them from the actual relative flow field.

Optimization Constraint

Based on equations (19, 20), the error function to be minimized over the set of relative flow vectors corresponding to a single segment or a possibly merged set of them is

$$E(\Omega_X, \Omega_Y, T_Z) = \sum_{i=1}^{n} W_i [\left(\frac{\Delta\alpha}{\delta}\right)_i - \Omega_Y x_i + \Omega_X y_i + T_Z / Z_i]^2. \tag{34}$$

where $T = (T_X, T_Y, T_Z)$ is the translation vector, $\Omega = (\Omega_X, \Omega_Y, \Omega_Z)$ is the rotation vector and Z_i is the depth of the environmental point corresponding to the image point (x_i, y_i). W_i is the weight (or confidence measure [4,5]) associated with the flow vector at the point (x_i, y_i). It is required to obtain the three MID parameters, $(\Omega_X, \Omega_Y, T_Z)$, that minimize the above functional. This is done by differentiating the functional with respect to each of them, and setting the resulting differential equal to zero. We can then solve the resulting matrix equation for the three MID parameters.

The computation at this step is fast since we only need to solve a 3×3 linear system, with some additional multiplication for the weighting process and summation over the set of relative flow vectors. For the purposes of the present formulation, we use the disparity field, δ, to provide the absolute depth information, Z_i. Thus, we obtain the absolute translation in depth, T_Z, and the two rotations in depth, Ω_X and Ω_Y.

Given a set of n flow vectors, let the solution to the minimization constraint be given by $\mathcal{P}^* = (\Omega_X^*, \Omega_Y^*, T_Z^*)$. The normalized value of the deviation in the actual relative flow field from that predicted by the solution $\mathcal{P}^*$ can then be given by

$$\sigma = \sqrt{\frac{E(\mathcal{P}^*)}{\sum_{i=1}^{n} W_i}}. \tag{35}$$

5. Simulation Results

In this section, we use stereoscopic motion to compute the MID parameters of the objects in the environment and of the camera. The results are shown for a set of simulations that were devised to cover the following categories of motion for rigid objects –

- translation in depth alone,
- rotation in depth alone,
- combined translation and rotation in depth, and
- general, independent object and camera motion.

The input data are the simulated flow fields corresponding to the left and right cameras as well as the simulated static disparity field between the two cameras for the first time instance. Note that the flow fields are dense and could only correspond to highly textured surfaces. In the first four experiments, we have shown the algorithm performance for *ideal* flow and disparity fields. In the fifth and sixth experiments, we show the algorithm performance with gaussian noise added to the flow fields. Note that the algorithm assumes as input, the optic flow and disparity values, and hence bypasses the correspondence problem.

Values of the translation parameters and the surface parameters are in focal units, the flow and disparity vectors are in pixel units, and the rotation parameters are in radians. The image is 128×128. The field of view of each of the cameras is 45°. The baseline is 0.5 focal units, giving rise to disparity values of about 4 to 6 pixels for the simulated environment. The size and position of

size	*position*	*Object Translation (focal units)*	
		Input	*Computed*
2,2,2	-3,-1,15	$T_Z = 1.00$	$T_Z^{comp} = 1.08$

Table 1: Translation in depth of sphere of radius=2. Camera is stationary.

the objects are given in focal units and are with respect to the left camera, coordinate system, as are the motion values.

Experiment 1 : Object Translating in Depth

This experiment demonstrates the chief advantage of using stereoscopic constraints in motion computation, i.e., the *fast* computation of the motion of objects translating in depth. Such motion represents the direct motion of the object along a line parallel to the line of sight.

The simulated input motion and the computed translation in depth for the moving object are shown in Table 1. The environment consists of two distinct surfaces :

1. a plane described by the equation $Z = 100$,

2. a sphere of radius=2, at position = $(-3, -1, 15)$ translating with $T_Z = 1.00$.

The camera system is stationary.

The simulated flow fields corresponding to the left and right cameras as well as the simulated disparity field between them at the first time instance are shown in Figures 3a, 3b, and 3c, respectively. Figure 3d represents the segmentation mask obtained from the segmentation of the left monocular optic flow field using Adiv's algorithm [1]. The translation in depth of the object is computed using this segmentation on the relative flow field (see Table 1).

Note that the computed value of the translation in depth is the absolute (not relative) value since disparity information is used to obtain the depth.

Experiment 2 : Object rotating in depth.

In this experiment, we demonstrate the ability of the model to find the rotation in depth of objects. This will be used in future studies to model the perception of the structure of rotating objects [10,29,30]. The motion of interest is a spinning motion about axes parallel to the image planes.

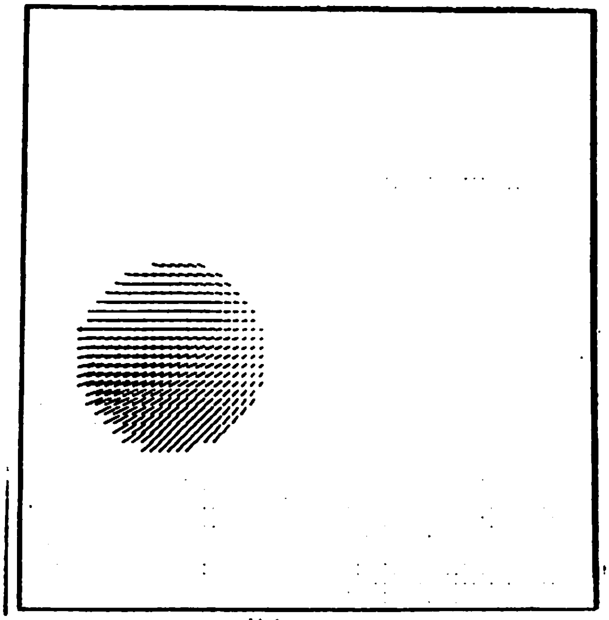

Figure 3a: Simulated, ideal, dense optic flow field for the left camera.

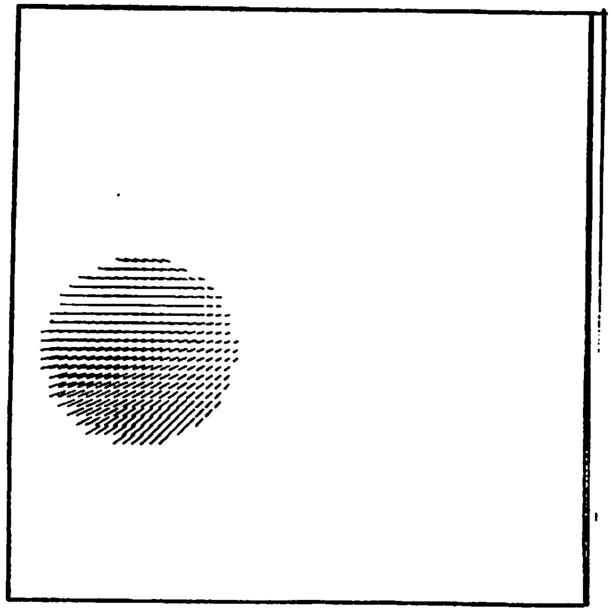

Figure 3b: Simulated, ideal, dense optic flow field for the right camera.

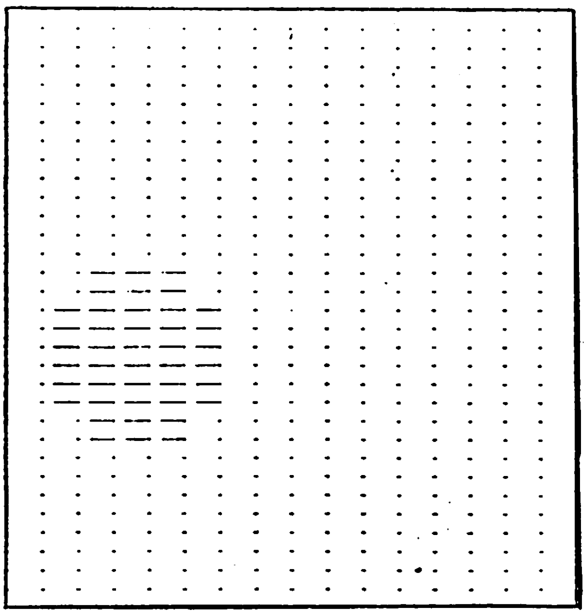

Figure 3c: Simulated, ideal, dense disparity field.

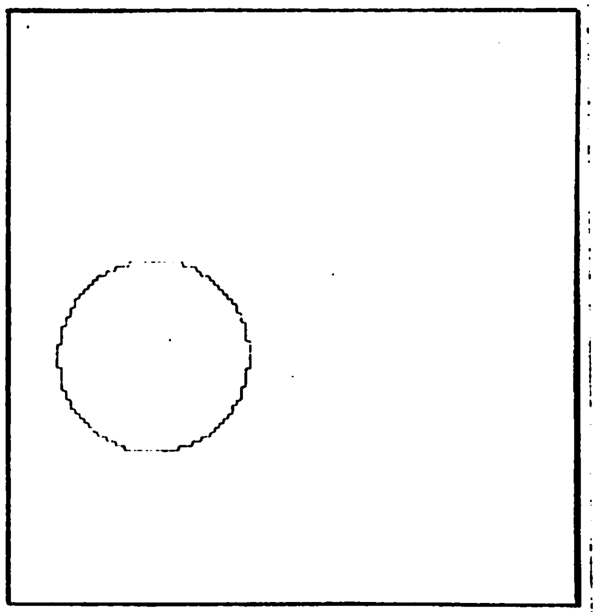

Figure 3d: Result of segmentation using Adiv's algorithm [1] on the left flow field.

size	*position*	*Object Translation (focal units)*	
		Input	*Computed*
2,2,2	-3,-1,15	$T_Z = 1.00$	$T_Z^{comp} = 1.08$

Table 1: Translation in depth of sphere of radius=2. Camera is stationary.

size	*position*	*Object Rotation (radians)*	
		Input	*Computed*
2,2,2	-3,-1,15	$\Omega_X = 0.05$	$\Omega_X^{comp} = 0.08$
		$\Omega_Y = 0.05$	$\Omega_Y^{comp} = 0.04$

Table 2: Rotation in depth of sphere of radius=2. Camera is stationary.

The simulated input motion and the computed rotation in depth for the moving object are shown in Table 2. The environment consists of two distinct surfaces :

1. a plane described by the equation $Z = 100$,

2. a sphere of radius=2, at position = $(-3, -1, 15)$ rotating with $\Omega_X = 0.05$ and $\Omega_Y = 0.05$.

The value of the rotation in depth parameters computed by the algorithm are shown in Table 2.

Experiment 3 : General motion in depth of object

It is shown here that the algorithm makes no assumptions about the motion of the object(s) and is applicable in the presence of both rotation and translation in depth. This example is a simulation of the motion of a spinning ball being thrown towards the observer. The object motion consists of translational components along the the X and Z axes, and a single rotational component about the X axis (both stated with respect to the left camera coordinate system). The camera system is stationary.

The values of both the translation in depth and the rotations in depth are computed and are shown in Table 3.

The input being simulated is as follows – The environment consists of two distinct surfaces :

1. a plane described by the equation $Z = 100$,

2. a sphere of radius=2, with position = $(1, 4, 15)$, with translation $(T_X = 0.5, T_Z = 1.20)$ and rotation $(\Omega_X = 0.05)$.

size	*position*	*Object Translation (focal units)*		*Object Rotation (radians)*	
		Input	*Computed*	*Input*	*Computed*
2,2,2	-3,-1,15	$T_X = 0.50$ $T_Z = 1.20$	$T_Z^{comp} = 1.12$	$\Omega_X = 0.05$	$\Omega_X^{comp} = 0.08$

Table 3: Motion in depth of sphere of radius=2. Camera is stationary. Note that only the MID parameters are computed.

Note in Table 3 that T_X is not computed. This is because our algorithm computes only the MID parameters.

Experiment 4 : General camera motion and independent object motion

In this example, we show the performance on a simulation of both camera motion and independent object motion. The camera motion is completely general with all the components of translation and rotation being present. The object motion has NO MID components, but has the other three components of motion, i.e., translations along the X and Y axes, and rotation about the Z axis.

The input being simulated is as follows - The environment consists of two distinct surfaces :

- the background described by a plane, $Z = X + 0.5Y + 50$,
- a sphere with position = $(9, 9, 30)$, radius = 2 and motion described by (T_X, T_Y, T_Z) = $(0.5, -0.5, 0.0)$ and $(\Omega_X, \Omega_Y, \Omega_Z) = (0.00, 0.00, -0.19)$.
- a *stationary* ellipsoid with position = $(-3, -1, 20)$ and size = $(2, 5, 2)$.

The motion of the camera is $(T_X, T_Y, T_Z) = (0.5, 0.5, 1.0)$ and $(\Omega_X, \Omega_Y, \Omega_Z) = (0.02, -0.02, 0.04)$.

The flow fields corresponding to the left and right cameras as well as the disparity field between them at the first time instance are shown in Figures 4a, 4b and 4c, respectively. Figure 4d represents the segmentation mask obtained from the left optic flow field to be used as a mask in the optimization stage.

Note that in Figure 4e, the minimization step of the algorithm results in the *stationary* ellipsoid getting merged with the background. This is because both have the same MID attributes relative to the moving camera. The independently moving sphere is still retained as a seperate region since it has MID attributes relative to the moving camera that are distinct from those of the stationary

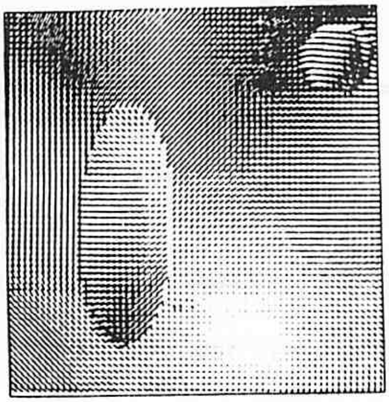

Figure 4a: Simulated *ideal*, dense optic flow field for the left camera.

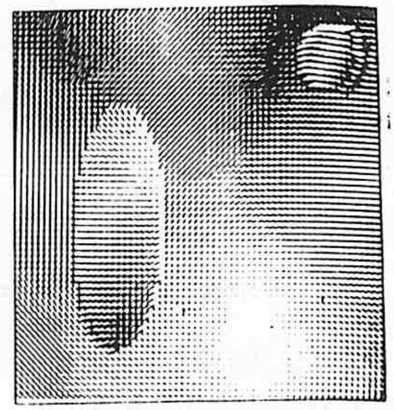

Figure 4b: Simulated *ideal*, dense optic flow field for the right camera.

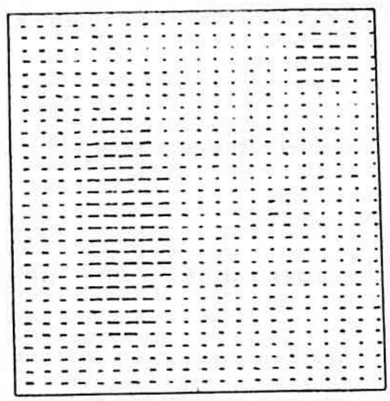

Figure 4c: Simulated *ideal* dense field of disparity vectors. Baseline is 0.5 focal units.

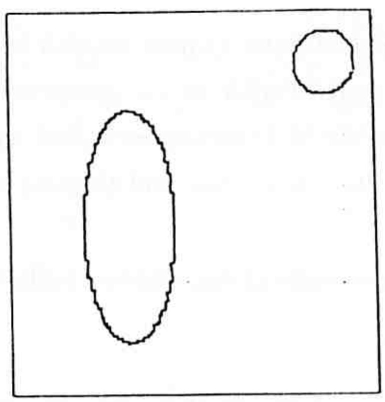

Figure 4d: Result of segmentation performed using Adiv's algorithm [1].

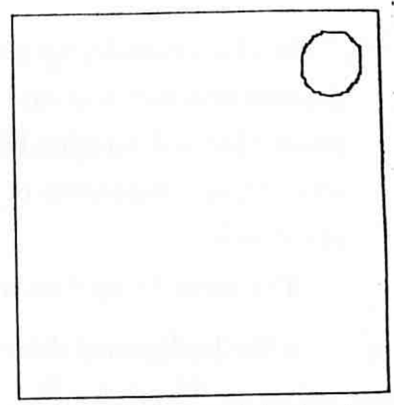

Figure 4e: Result of merger in the optimization step of the algorithm. Note that the independently moving sphere is picked out.

	size	*position*	*Object Translation (focal units)*		*Object Rotation (radians)*	
			Input	*Computed*	*Input*	*Computed*
sphere	2,2,2	9,9,30	$T_X = 0.50$ $T_Y = -0.5$ $T_Z = 0.00$	$T_Z^{comp} = 0.11$	$\Omega_X = 0.00$ $\Omega_Y = 0.00$ $\Omega_Z = -0.19$	$\Omega_X^{comp} = 0.04$ $\Omega_Y^{comp} = 0.02$
	size	*position*	*Object Translation (focal units)*		*Object Rotation (radians)*	
ellipsoid	2,5,2	-3,-1,20	stationary			
plane	Z = X + 0.5Y + 50		stationary			

	Camera Translation (focal units)		*Camera Rotation (radians)*	
	Input	*Computed*	*Input*	*Computed*
camera	$T_X = 0.50$ $T_Y = 0.05$ $T_Z = 1.0$	$T_Z^{comp} = 1.2$	$\Omega_X = 0.02$ $\Omega_Y = -0.02$ $\Omega_Z = 0.04$	$\Omega_X^{comp} = 0.03$ $\Omega_Y^{comp} = -0.01$

Table 4: General camera motion with independent object motion.

<table>
<tr><td></td><td>size</td><td>position</td><td colspan="2">Object Translation (focal units)</td><td colspan="2">Object Rotation (radians)</td></tr>
<tr><td></td><td></td><td></td><td>Input</td><td>Computed</td><td>Input</td><td>Computed</td></tr>
<tr><td>sphere</td><td>2,2,2</td><td>9,9,30</td><td>$T_X = 0.50$
$T_Y = -0.5$
$T_Z = 0.00$</td><td>$T_Z^{comp} = 0.11$</td><td>$\Omega_X = 0.00$
$\Omega_Y = 0.00$
$\Omega_Z = -0.19$</td><td>$\Omega_X^{comp} = 0.04$
$\Omega_Y^{comp} = 0.02$</td></tr>
<tr><td></td><td>size</td><td>position</td><td colspan="2">Object Translation (focal units)</td><td colspan="2">Object Rotation (radians)</td></tr>
<tr><td>ellipsoid</td><td>2,5,2</td><td>-3,-1,20</td><td colspan="4">stationary</td></tr>
<tr><td>plane</td><td colspan="2">Z = X + 0.5Y + 50</td><td colspan="4">stationary</td></tr>
<tr><td></td><td colspan="3">Camera Translation (focal units)</td><td colspan="3">Camera Rotation (radians)</td></tr>
<tr><td></td><td>Input</td><td colspan="2">Computed</td><td>Input</td><td colspan="2">Computed</td></tr>
<tr><td>camera</td><td>$T_X = 0.50$
$T_Y = 0.05$
$T_Z = 1.0$</td><td colspan="2">$T_Z^{comp} = 1.2$</td><td>$\Omega_X = 0.02$
$\Omega_Y = -0.02$
$\Omega_Z = 0.04$</td><td colspan="2">$\Omega_X^{comp} = 0.03$
$\Omega_Y^{comp} = -0.01$</td></tr>
</table>

Table 4: General camera motion with independent object motion.

background. This is an example of the fact that the algorithm does not pick out object masks, but does pick out regions in the image with the same set of MID parameters.

We note that the computed values of the MID parameters for the independently moving sphere are inaccurate. We attribute this to the following fact that the number of relative flow vectors (19,20) in the mask corresponding to the segmentation of the sphere is small. When the values are computed using the least squares minimization process on this small set of vectors, errors can be expected.

Experiment 5 : Object Translating in Depth – Noisy Optic Flow Fields

In this experiment, we show the performance of the algorithm for the computation of the translation in depth of the object with *noisy optic flow fields* as input.

The environment simulated is identical to that in Experiment 1, with the camera being stationary. We have added gaussian noise of $\sigma = 0.3$ to the optic flow fields for the left and the right camera. These are shown in Figures 5a and 5b respectively. The results of the segmentation on the left optic flow field is shown in Figure 5c. Table 5 shows the results of the computation of the translation in depth for the object.

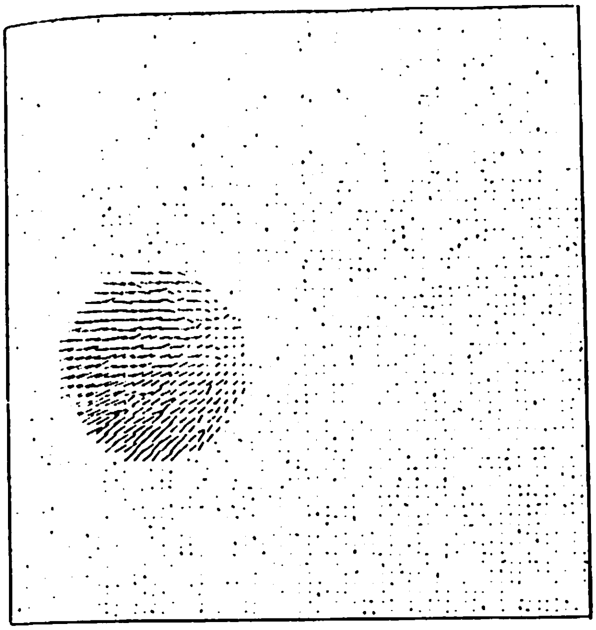

Figure 5a: Simulated *noisy*, dense optic flow field for the left camera.

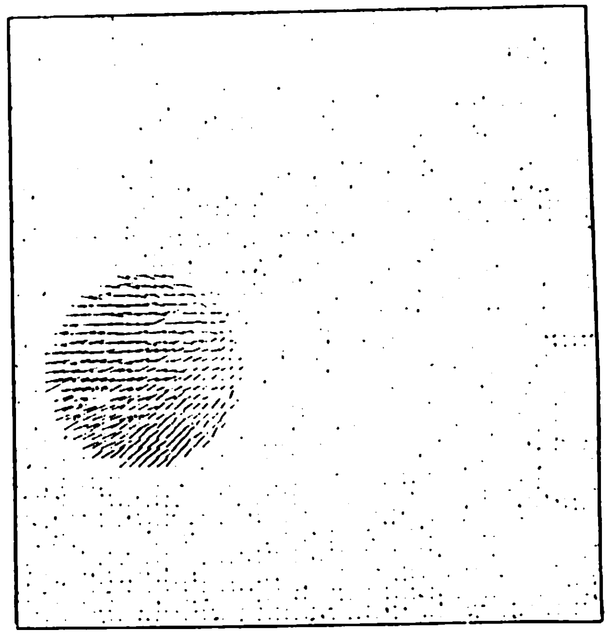

Figure 5b: Simulated *noisy*, dense optic flow field for the right camera.

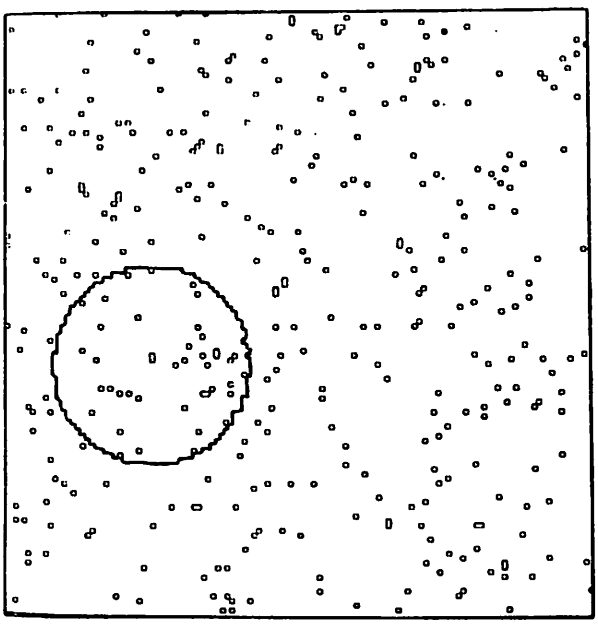

Figure 5c: Result of segmentation performed using Adiv's algorithm [1].

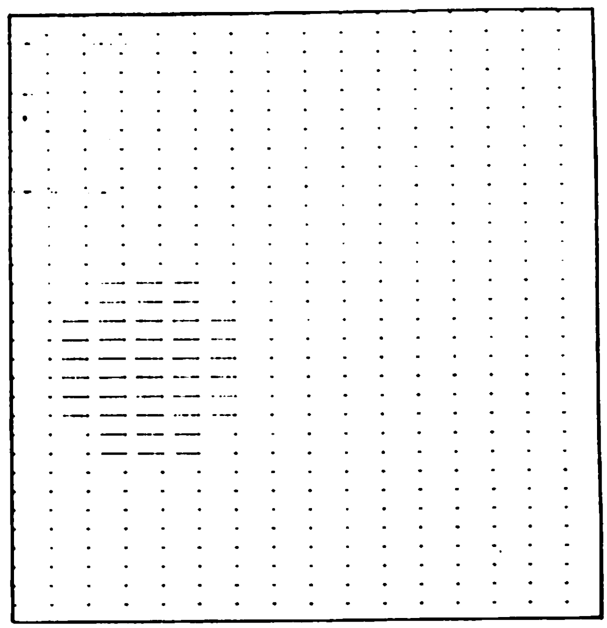

Figure 5d: Simulated *ideal* dense field of disparity vectors. Baseline is 0.5 focal units.

size	*position*	*Object Translation (focal units)*	
		Input	*Computed*
2,2,2	-3,-1,15	$T_Z = 1.00$	$T_Z^{comp} = 1.10$

Table 5: Translation in depth of sphere of radius=2. Gaussian noise of $\sigma = 0.3$ added to the optic flow fields. Camera is stationary.

size	*position*	*Object Translation (focal units)*	
		Input	*Computed*
2,2,2	-3,-1,15	$T_Z = 1.00$	$T_Z^{comp} = 1.10$

Table 5: Translation in depth of sphere of radius=2. Gaussian noise of $\sigma = 0.3$ added to the optic flow fields. Camera is stationary.

Note that the error in the computation is not significantly greater than that shown in Table 1 for experiment 1.

Experiment 6 : General camera motion and independent object motion – Noisy Optic Flow Fields

In this example, we show the performance of the algorithm in the presence of both camera motion and independent object motion with ***noisy optic flow fields*** as input.

The camera motion is completely general with all the components of translation and rotation being present. The object motion has NO MID components, but has the other three components of motion, i.e., translations along the X and Y axes, and rotation about the Z axis.

Note that the simulated input motion and environment description is identical to that in Experiment 4. However, gaussian noise of $\sigma = 0.3$ is added to the corresponding optic flow fields for the left and the right cameras. These are shown in Figures 6a and 6b. Figure 6c shows the results of segmenting the left optic flow field using Adiv's technique [1]. In Table 6, we show the MID parameters computed for the camera and the independently moving sphere.

Note that in Figure 6d, the ***stationary*** ellipsoid is still getting merged with the background, while the independently moving sphere continues to be picked out, just as in Experiment 4 (see Figure 4e).

The MID parameters computed for the independently moving sphere are even more inaccurate than those computed for the ideal flow fields in Experiment 4. We attribute this to the fact that the number of relative flow vectors (19,20) in the mask corresponding to the segmentation of the sphere is small as well as the fact that the error in the relative flow vectors increases with non-ideal flow fields. We note that the computed values of the MID parameters for the camera are identical to those values computed for the ideal flow fields in Experiment 4 (see Table 4), demonstrating that the optimization step shows robust performance when the number of relative flow vectors available

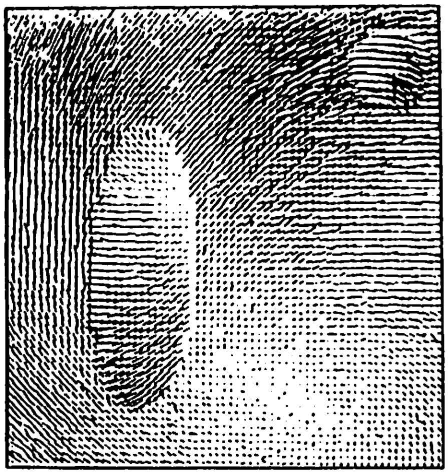

Figure 6a: Simulated *noisy*, dense optic flow field for the left camera.

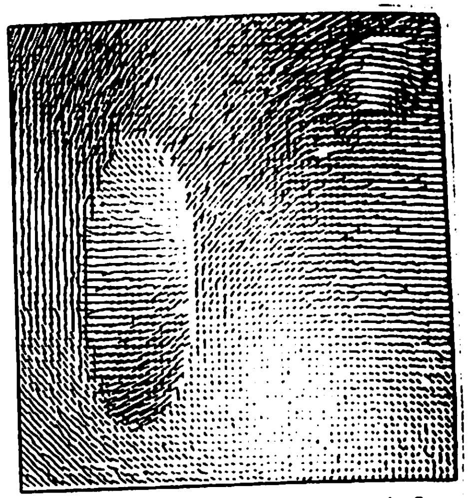

Figure 6b: Simulated *noisy*, dense optic flow field for the right camera.

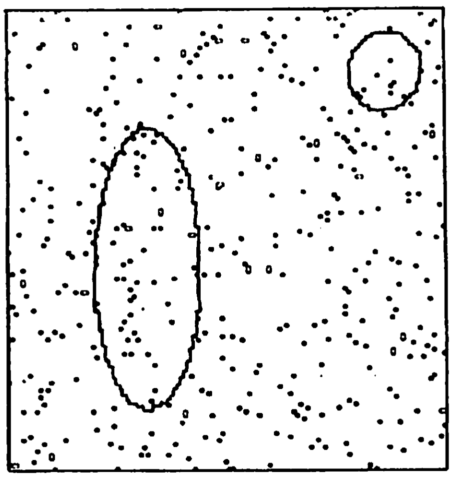

Figure 6c: Result of segmentation performed using Adiv's algorithm [1].

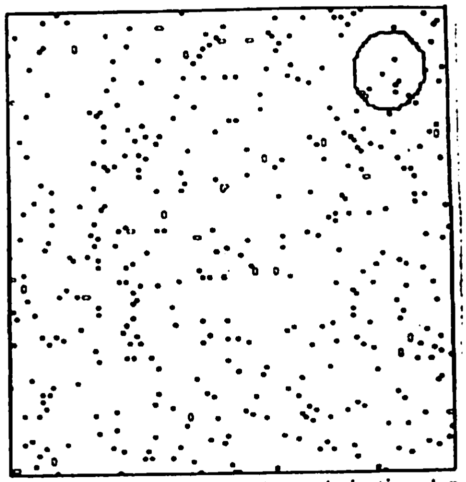

Figure 6d: Result of merger in the optimization step of the algorithm. Note that the independently moving sphere is still picked out while the stationary ellipsoid is merged with the background.

	size	*position*	*Object Translation (focal units)*		*Object Rotation (radians)*	
			Input	*Computed*	*Input*	*Computed*
sphere	2,2,2	9,9,30	$T_X = 0.50$ $T_Y = -0.5$ $T_Z = 0.00$	$T_Z^{comp} = 0.46$	$\Omega_X = 0.00$ $\Omega_Y = 0.00$ $\Omega_Z = -0.19$	$\Omega_X^{comp} = 0.08$ $\Omega_Y^{comp} = 0.06$
	size	*position*	*Object Translation (focal units)*		*Object Rotation (radians)*	
ellipsoid	2,5,2	-3,-1,20	stationary			
plane	Z = X + 0.5Y + 50		stationary			
	Camera Translation (focal units)		*Camera Rotation (radians)*			
	Input	*Computed*	*Input*	*Computed*		
camera	$T_X = 0.50$ $T_Y = 0.05$ $T_Z = 1.0$	$T_Z^{comp} = 1.2$	$\Omega_X = 0.02$ $\Omega_Y = -0.02$ $\Omega_Z = 0.04$	$\Omega_X^{comp} = 0.03$ $\Omega_Y^{comp} = -0.01$		

Table 6: General camera motion with independent object motion. Gaussian noise of $\sigma = 0.3$ added to the optic flow fields.

	size	position	Object Translation (focal units)		Object Rotation (radians)	
			Input	Computed	Input	Computed
sphere	2,2,2	9,9,30	$T_X = 0.50$		$\Omega_X = 0.00$	$\Omega_X^{comp} = 0.08$
			$T_Y = -0.5$		$\Omega_Y = 0.00$	$\Omega_Y^{comp} = 0.06$
			$T_Z = 0.00$	$T_Z^{comp} = 0.46$	$\Omega_Z = -0.19$	
	size	position	Object Translation (focal units)		Object Rotation (radians)	
ellipsoid	2,5,2	-3,-1,20	stationary			
plane	Z = X + 0.5Y + 50		stationary			
	Camera Translation (focal units)		Camera Rotation (radians)			
	Input	Computed	Input	Computed		
camera	$T_X = 0.50$		$\Omega_X = 0.02$	$\Omega_X^{comp} = 0.03$		
	$T_Y = 0.05$		$\Omega_Y = -0.02$	$\Omega_Y^{comp} = -0.01$		
	$T_Z = 1.0$	$T_Z^{comp} = 1.2$	$\Omega_Z = 0.04$			

Table 6: General camera motion with independent object motion. Gaussian noise of $\sigma = 0.3$ added to the optic flow fields.

for the computation is not small, even if they are erroneous.

6. Summary

We have so far described how stereoscopic motion can be used to extract the MID parameters. The chief motivation was the need to obtain the MID parameters, particularly translation in depth, as quickly as possible. These parameters can be obtained from a general purpose motion algorithm such as Adiv's [1], which computes all the six motion parameters. But these algorithms are computationally intensive and do not permit a *quick yet reliable* grouping of the information from the imagery into regions corresponding to objects moving in depth. One of the chief reasons for this is that the computations need to deal with nonlinear equations that relate the flow information to the motion and structure parameters.

Use of stereoscopic motion for the computation of the MID parameters simplifies and speeds up the computation because of the following reasons -

1. The chief advantage provided by the integration of stereo and motion information for extracting motion in depth is that stereo information provides additional constraints that make it possible to formulate a *linear* relationship between the data in the flow and disparity fields

and the motion in depth parameters, {see equations (19, 20)}. This, in turn, makes it possible to devise a direct computation without any hypothesizing of the motion parameters such as is done in [1,21].

2. The technique does *not* consider all possible subsets of the segments for grouping purposes. It only considers the *neighbouring* segments for merger decisions. This is reasonable since we are searching for groups in the image that correspond to the same three MID parameters, rather than seeking object masks. This makes the complexity of this step linear, rather than exponential (as in [1]).

3. The computation of the three MID parameters are the result of a direct computation using the optimization constraint on the relative flow field. This can be used to directly compute the remaining three motion parameters, again, without employing a hypothesize and test scheme. Hence, extracting all the motion parameters becomes a problem of handling two sets of linear functionals.

4. The algorithm can be employed on motion sequences that include several independently moving objects and involve the general translation and rotation of the objects and the camera.

A problem that we hope to resolve in future work has to do with the nature of the relative flow fields. Since they are computed as a difference of the two flow fields, they are susceptible to error if the flow values are small. Hence, the optimization step is more accurate with larger motion values. Incorporating minimization norms that take this into account may help.

In the current implementation, we use the disparity field to provide us with the depth information in the computation. In future work, we hope to use the current results as a startup process and extend to a multi-frame paradigm for computing the motion over several frames as well as using this disparity information as a prediction of the depth and refining the depth map over several frames.

Acknowledgements

The authors would like to thank P. Anandan, G. Adiv, Debbi Strahman and Philip Kahn for many valuable discussions. Thanks are also due to Ed Riseman for his comments on the manuscript, and to the developers of the VISIONS system - in particular, Robert Heller - for always being helpful in ironing out implementation glitches.

REFERENCES

[1] G. Adiv," Determining Three-Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence* Vol. PAMI-7, July 1985, 384–401.

[2] G. Adiv,"Interpreting Optical Flow," *Ph. D. Dissertation*, 1985, University Of Massachusetts, COINS TR 85-35.

[3] J. (Y). Aloimonos and I. Rigoutsos," Determining the 3-D motion of a rigid planar patch without correspondence, under perspective projection," *Proceedings of the IEEE Motion Workshop, Charleston, S.C.*, 1986, 167–174.

[4] P. Anandan, "Computing Dense Displacement Fields with Confidence Measures In Scenes Containing Occlusion," *SPIE Intelligent Robots and Computer Vision Conference* Vol. 521, 1984, 184–194. See also COINS TR 87-2, (UMass, Amherst).

[5] P. Anandan and R. Weiss, "Introducing A Smoothness Constraint In A Matching Approach For The Computation Of Optical Flow Fields," *IEEE Third Workshop On Computer Vision: Representation And Control* 1985, 186–194.

[6] P. Balasubramanyam, "Computation Of Motion-in-Depth Parameters Using Stereoscopic Motion Constraints " *IEEE Computer Society Workshop on Computer Vision* 1987, 349–351.

[7] P. Balasubramanyam and M. A. Snyder, "Computation Of Motion-in-Depth Parameters: A First Step In Stereoscopic Motion Interpretation " *Proceedings Of The DARPA Image Understanding Workshop* April 1988, 907–919.

[8] J. Barron, "A Survey Of Approaches For Determining Optic Flow, Environmental Layout and Egomotion," University Of Toronto Technical Report No. RBCV-TR-84-5, 1984.

[9] K. I. Beverley and D. Regan, "Evidence for the Existence of Neural Mechanisms Selectively Sensitive to the Direction of Movement In Space," *Journal Of Physiology*, **235**, 1973, 17–29

[10] M. L. Braunstein, "Perception of Rotation in Depth: The Psychological Evidence," *ACM Workshop on Motion* 1983, 119–124.

[11] J. Q. Fang and T. S. Huang, "Estimating 3-D Movement of a Rigid Object: Experimental Results," *Proceedings of the International Joint Conference on Artificial Intelligence, Karlsruhe, Germany* 1983, 1035–1037.

[12] B. K. P. Horn and E. J. Weldon, "Computationally Efficient Methods For Recovering Translational Motion," *Proceedings of the First International Conference On Computer Vision* June, 1987, 2–11.

[13] T. S. Huang and S. D. Blostein, "Robust Methods for Motion Estimation based on Two Sequential Stereo Image Pairs," *Proceedings of the IEEE Motion Workshop, Bellaire, Mich.*, 1985, 518–523.

[14] M. R. M. Jenkin, "The Stereopsis of Time-Varying Images," *Technical Report*, University of Toronto, Toronto, Ontario, Canada, (RBCV-TR-84-3), Sept. 1984.

[15] J. S. Lappin, J. F. Doner and B. L. Kottas, "Minimal Conditions for the Visual Detection of Structure and Motion in Three Dimensions," *Science*, 1980, **209**, 717–719.

[16] D. T. Lawton, "Processing Dynamic Image Sequences from a Moving Sensor," *Ph.D Dissertation* (TR 84-05), Computer and Information Science Dept., University Of Massachusetts, 1984.

[17] H. C. Longuet-Higgins and K. Pradzny, "The Interpretation of a Moving Retinal Image," *Proceedings of the Royal Society of London*, July 1980, **B**(208):385–397.

[18] W. R. Miles, "Movement interpretations of the silhouette of a revolving fan," *American Journal of Psychology*, 1931, **43**, 392–405.

[19] K. M. Mutch, "Determining Object Translation Information using Stereoscopic Motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. PAMI-8, Nov. 1986, 750–755.

[20] J. T. Petersik, "Rotation Judgements and Depth Judgements: Seperate or Dependent Processes," *Perception and Psychophysics*, 1980, **27**, 588–590.

[21] K. Pradzny, "Determining the Instantaneous Direction of Motion from Optical Flow Generated by a Curvilinearly Moving Observer," *Proceedings of the IEEE Conference on Pattern Recognition and Image Processing, Dallas, Texas* 1981, 109–14.

[22] D. Regan and K. I. Beverley, "Binocular and Monocular Stimuli for Motion-In-Depth: Changing Disparity and Changing Size Inputs Feed the Same Motion-In-Depth Stage," *Vision Research*, 1979, **19**, 1331–1342

[23] D. Regan, "Visual Processing of Four Kinds of Relative Motion," *Vision Research*, 1986, **26**, 127–145.

[24] W. Richards, "Structure from Motion and Stereo," *Journal Of the Optical Society of America* Feb. 1985, **2** 343–349.

[25] J. H. Rieger and D. T. Lawton, "Determining the Instantaneous Axis of Translation from Optic Flow Generated by Arbitrary Sensor Motion," *Proceedings of the Workshop on Motion: Representation and Perception, Toronto, Canada* 1983, 33–41.

[26] J. W. Roach and J. K. Aggarwal, "Determining the Movement Of Objects from a Sequence of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**, Nov. 1980, 554–562.

[27] W. B. Thompson, K. M. Mutch and V. A. Berzins, " Analyzing Object Motion Based on Optical Flow," *Proceedings of the International Conference on Pattern Recognition, Montreal, Canada* 1984, 791-794.

[28] R. Y. Tsai and T. S. Huang, "Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces," *IEEE Transactions on Pattern Analysis and Machine Inteligence* **6** Jan. 1984, 13–27.

[29] S. Ullman, The Interpretation Of Visual Motion, MIT Press, 1979, Cambridge & London.

[30] S. Ullman, "Maximizing Rigidity: The Incremental Recovery Of 3-D Structure From Rigid And Rubbery Motion," *A.I.Memo No. 721*, MIT, June, 1983.

[31] A. M. Waxman and J. H. Duncan, "Binocular Image Flows: Steps toward stereo-motion fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. PAMI-8, Nov.1986, 715–729.

[32] A. M. Waxman and K. Wohn. Image Flow Theory: A Framework for 3-D Inference from Time-Varying Imagery, *Chapter in Advances in Computer Vision (Erlbaum Publishers)*.