A Review of Motion and Stereopsis Research *

P. Anandan

COINS Technical Report 85-52**
December, 1985

Abstract

This report contains a review of some of the major approaches described in the computer vision literature for the analysis of motion and stereopsis. The report is intended to introduce a researcher in the field of computer vision to the state of the art in these areas. The major research approaches have been organized and classified in a coherent manner and the basic ideas contained in them are described. This is not intended to be an exhaustive survey, although an attempt has been made to refer to most of the relevant techniques found in the literature.

---

* *to appear (in Spanish) as a chapter in* **Vision por Computador**, ed. Carme Torras, *to be published by Alianza Editorial, Spain.*

** The report is sponsored by Allen R. Hanson and Edward M. Riseman. This research was supported by DARPA under grant N00014-82-K-0464 and NSF under grant DCR-8318776.

Contents

1 INTRODUCTION 3

1.1 Stereopsis 4

1.2 Motion 5

1.3 How to read this chapter 7

2 THE CORRESPONDENCE PROBLEM 8

2.1 Intensity based correspondence 10

2.1.1 Gradient based schemes 10

2.1.2 Correlation matching schemes 13

2.1.3 A few remarks 16

2.2 Token matching schemes 16

2.2.1 Point matching schemes 18

2.2.2 Matching edge tokens 23

2.3 The use of non-local constraints 25

2.4 Spatial frequency analysis 29

3 DETERMINING 3-D MOTION AND STRUCTURE 32

3.1 The geometry of flow fields 33

3.2 Processing a restricted class of motions 37

3.3 Approaches based on point correspondences 41

3.4 Techniques requiring optic flow 45

3.4.1 Local techniques 46

3.4.2 Global techniques 50

3.5 Concluding remarks 52

4 STEREOPSIS **53**

4.1 The geometry of stereopsis 54

4.2 Geometric and physical constraints 55

4.2.1 The use of physical constraints 57

4.3 Using domain specific constraints 58

4.3.1 The approach used in the 3-d Mosaic system 58

4.3.2 The rule based analysis at Stanford 59

4.4 Stereopsis algorithms based on human vision 60

4.5 Obtaining 3-d descriptions 64

4.6 A few remarks 65

5 CONCLUSION **65**

5.1 Integration of motion and stereopsis 67

5.2 Processing longer image sequences 68

1 INTRODUCTION

There are a number of techniques that provide the information necessary to obtain the 3 dimensional structure of the environment from a single static visual image, such as shape from shading, deformation of areas, and vanishing point analysis. However, these techniques are not always reliable. They may fail under unfavourable illumination conditions or when the underlying assumptions regarding the shape of the world surfaces are invalid. If, however, two cameras located a short distance apart are used, the two distinct views provided by them can be combined to produce reliable 3-d information about the environment.

In a similar vein, one of the key features of an object that usually distinguishes it from other objects in the environment is its movement relative to them. Even when an object is camouflaged by its similarity in appearance to other objects, any independent movement of the object immediately gives it away. In addition, if there is a relative movement between

Thanks are due to Mark Snyder whose detailed comments have made this somewhat more comprehensible, to Lance Williams who contributed section 4.4, and to Poornima Balasubramanyam for her help in section 3.4.2.

the camera and the object, the viewer is automatically provided with several distinct views of the object, and therefore with 3-d information. In general, use of the dynamic properties of the objects in the images can provide reliable information about segementation of the image into distinct objects, their 3-d structures, and their dynamic characteristics.

The two most common methods of obtaining two images from two distinct views are **stereopsis** and **motion**. Stereopsis is when two images are obtained simultaneously by two cameras. Motion is when several images are taken one after another by a single camera and in the meanwhile, there is relative movement between the camera and the environment.

In this chapter, we explain the stereopsis and motion approaches to obtaining the 3 dimensional structure of the environment, and outline some of the major efforts described in the literature. We further consider the geometric "constraints" involved and the issues involved in applying these constraints to successfully compute the 3-d structure and movement.

1.1 Stereopsis

As mentioned before, stereopsis refers to the situation when two images are obtained simultaneously from two distinct view points. In most applications of stereopsis, it is common to orient the cameras such that their image planes are perpendicular to the ground plane and their optical axes are parallel to each other. Usually the displacement between the camera locations is horizontal and parallel to the image plane.

Given the two images, the task at hand is to combine them to provide 3-d information about the objects in the image. All the approaches described in this chapter assume that stereo analysis proceeds without the aid of other processes, such as texture, region,

and shape analyses. The process usually consists of two stages - the establishment of the *correspondence* between the points in the two images to provide a *disparity* and then a *depth* map, followed by some process that uses the depth information to discover and describe the surfaces in the 3-d environment.

Before we proceed further, we define a few key terms. The *correspondence problem* is the task of identifying events in the two images as images of the same event in the 3-d environment. The *disparity* is the distance between the locations in the two images of the two corresponding events. When the optical axes are parallel to each other, the *depth* of a point is the distance along the optical axis from the image planes.

Finally, a key concept in stereopsis is *vergence*. In biological vision, vergence is the process of converging the two eyes to fixate at points at different depths. In industrial machine vision, the optical axes of the cameras are maintained parallel to each other, and vergence is achieved by shifting the images relative to each other by different amounts.

1.2 Motion

Motion processing can be broadly divided into two categories: (1) the camera moves and the environment is stationary, and (2) there are independently moving objects in the scene. The first case is easier to analyze and process, as will be seen from the large number of techniques that have been developed for this purpose.

The most common approach taken towards motion analysis is one in which the processing proceeds bottom-up – similar to the approach mentioned for stereopsis. The movement of individual points in the images is computed first, followed by a process that determines the motion of the camera, as well as the location, 3-d structure, and motion of the objects

in the scene.

It must be noted, however, that not all researchers have adopted this approach. Some approaches attempt to simultaneously compute the movement of the individual points and the motion of the camera, while others attempt to first segment the image and then compute the motion of the segments.

One important term used in motion research is **optic flow**. Different authors have defined this differently. Following Lawton [Lawt84], optic flow can be broadly defined as the vector field representing the changes in the positions of the images of environmental points over time. The term was introduced by the psychologist J. J. Gibson, although Gibson did not deal with the computation of optic flow. The following quote from [Lawt84] demonstrates the ambiguity in the definition of this term.

> There is some ambiguity in the definition of optic flow in the literature (even with respect to the phrase itself, since *optical flow* or even *optic flows* are used). Some refer to the flow field as being entirely independent of images, and instead view it as a representation of the changes in environmental directions over time. To others it is a basic description of image motion determined from image intensity changes and not necessarily related to environmental motion... A further source of ambiguity is that some people refer to the optic flow as a continuous vector field in which the vectors are instantaneous velocity vectors, while others refer to it as a field of discrete displacement vectors.

Strictly speaking, it is necessary to distinguish between *optic flow*, which is the field of instantaneous 2-d velocity vectors of the points in the image on the image plane, and

displacement field, which is the field of discrete displacement vectors connecting the location of the same image-point in successive image frames. It must be noted, however, that when the time interval between the frames is small enough, the displacement field is a good approximation to the optic flow. This is the view point taken by many researchers. For simplicity, the term *optic flow* is used in this chapter both for "displacement field" and for "optic flow". The precise meaning will usually be evident from the context in which the term is used.

The usual approach to motion analysis consists of two steps – the computation of optic flow, followed by its interpretation to provide the 3-d structure and motion of the objects in the scene as well as the motion of the camera. The computation of optic flow is similar to the correspondence problem mentioned earlier in this section. In fact, it is common to regard the correspondence problem in stereopsis as a special case of motion correspondence. However, in stereopsis, the knowledge of the relative locations of the cameras constrains the search for corresponding points in a manner that is not possible in motion analysis.

Finally, we mention one important limitation of current approaches to motion analysis. Most of the techniques for motion analysis deal with only two frames. Some initial approaches to multi-frame analysis, as well as some speculative ideas, are described at the end of this chapter.

1.3 How to read this chapter

The rest of this chapter is divided into four sections. In section 2 we discuss various approaches used to solve the correspondence problem (or to compute optic flow), although vergence is not discussed. In section 3, we discuss methods that can be used to derive

surface from moiton with known optic flow, as well as some of the techniques that do not require optic flow. In section 4, we consider correspondence algorithms specific to stereopsis and some issues regarding the 3-d interpretation of the results of stereopsis. Finally, in section 5, we summarize the state of the art and describe some open issues and problems.

We should note that this chapter is not intended as a survey of the various techniques used by researchers. It should be regarded as an introductory review of stereopsis and motion research. We will explain the principles underlying the major types of methods studied by researchers, and will not focus on fine variations on the themes. Finally, at the end we will provide a bibliography.

2 THE CORRESPONDENCE PROBLEM

Identifying image "events" that correspond to each other is the primary task of both motion and stereo analysis. The term "events" is used here in a broad sense, to mean any identifiable structure in the image – e.g., image intensities in a neighborhood, edges, lines, texture markings, etc.

The techniques that rely on the similarity of the light intensity reflected from a scene location in the two frames as the basis for determining correspondence are called *intensity based* approaches. Methods that identify stable image structures, and use them as tokens for finding correspondences are referred to as *token based* approaches.

The most popular way of solving the correspondencce problem is to divide it into one or two parts. The first is the local correspondence problem, which provides partial or total constraints on the displacement of a point in the image, based on image information

in the immediate neighborhood of that point. Usually the local correspondence is solved (partially or fully) *independently* at all points of interest in the image. The second part, where used, consists in applying a non-local constraint on the flow field. This is usually an assumption of the spatial smoothness of the flow field, or one that is derived from the geometry of rigid bodies in motion. This constraint can be either global or semi-global, depending on whether or not explicit boundaries are recognized, across which the constraint is not allowed to propagate.

It is also possible to impose on top of this framework for the computation of displacement fields, a multi-frequency, multi-resolution approach. In this approach the images are pre-processed with a set of band-pass filters which are spatially local and which decompose the spatial frequency-spectrum in the image in a convenient way. The outputs from the corresponding filters applied to the two images are matched, and the matching results from the different filters at the same location in the image are combined using a consistency constraint.

We first consider the local correspondence problem, and then the use of a non-local constraint. Of the schemes that find the local correspondences, intensity based methods will be described first, followed bya description of methods that generate point tokens and match them, and methods that use linear structures in the image. Non-local constraints can be applied to almost any of these approaches to solve the local correspondence problem, although their precise algorithmic form will vary. Finally, the use of spatial frequency channels will be treated. The use of structured tokens for matching is not discussed here, since much of such work is preliminary and are rarely used. A list of papers describing such approaches is included in the reading list.

All of the approaches are described primarily from the viewpoint of motion analysis, although many of them are also applicable to stereopsis. The techniques that are specifically suited for stereopsis will be discussed in section 4

2.1 Intensity based correspondence

The most direct approach to correspondence is to match the light intensity refelected from a point in the environment and recorded in the two images. Assuming that the time difference between the generation of the images is small (in the case of stereo this is given to be zero), the intensity of the image of a specific environmental point is likely to be the same in both images. This constancy of the image intensity of a point across the images is usually called the *intensity-constancy constraint.*

Intensity based schemes are those that use this intensity constancy constraint. They can be broadly divided into two classes, *gradient-based schemes* and *correlation matching schemes.*

2.1.1 Gradient based schemes

Consider the simple situation when the points in an image are translating parallel to the image plane. Although this situation is rare in perspective images, it is convenient for explaining the gradient-based schemes. Let $I(x, y, t)$ be the intensity at a point (x, y) on the image plane at time t.

Assume that a point at location (x, y) in the image at time t moves to the location $(x+\delta x, y+\delta y)$ at time $t+\delta t$. The intensity constancy assumption states that the intensity

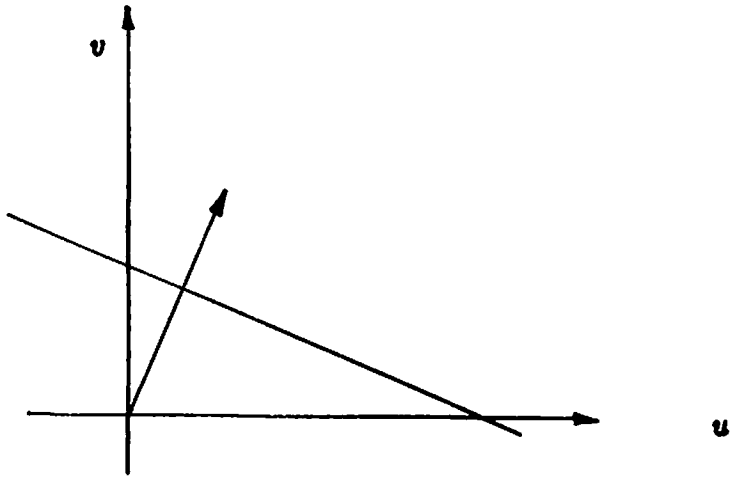

Figure 1: **The intensity constraint**

of this point is the same in the two images, i.e.,

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t)$$

Using this we find [Horn80]:

$$I_x u + I_y v + I_t = 0, \tag{1}$$

where $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, $I_t = \frac{\partial I}{\partial t}$, and u and v are the x and y components of the image-velocity of the point at time t.

In gradient based methods, equation (1) is the intensity constraint. This can also be represented graphically as the locus of all points in the (u, v) plane that satisfy the intensity constraint (see figure 1). Thus, the intensity constraint formulated here only *partially constrains* the image velocity at a point. The locus is a line perpendicular to the local image intensity gradient vector (I_x, I_y).

The intensity constraint can be written in the form of an error,

$$E_I = I_x u + I_y v + I_t$$

which is usually included in a minimization process, along with an error involving the global constraint on the displacement field.

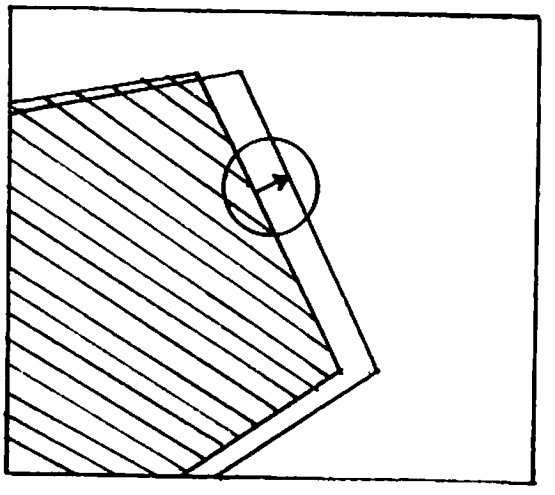

Figure 2: The aperture problem. Note that although the edge moves horizontally, if attention is focused to the circular area shown, the edge appears to move perpendicular to its orientation.

The aperture problem and the normal flow The true velocity vector at a point in the image can be written as a sum of the component parallel to, and the component perpendicular to, the intensity gradient vector ∇I. The intensity constraint gives only the component of image velocity in the direction parallel to the spatial intensity gradient vector. The component parallel to the intensity constraint line is undetermined.

The lack of any information about the component of velocity parallel to the intensity constraint line is known as the *aperture problem.* This term comes from psychophysical studies of biological vision, where it is well known that if attention is focused on a narrow area around a point on a line, the line appears to move perpendicular to itself. This is illustrated in figure 2.

The perceived velocity is thus the component of the true velocity in this direction.

Modifications to the intensity constraint The above formulation of the intensity constraint is due to Limb and Murphy [Limb75]. It has also been derived and used by Fennema and Thompson [Fenn79], and Horn and Schunck [Horn80]. There have been a number of other formulations since then, the most notable of which are those of Cornelius and Kanade [Corn83] and Nagel [Nage83a].

Cornelius and Kanade relax the assumption that the intensity of the point stays constant over time. They state that temporal changes in the intensity at a point (i.e, the total derivative $\frac{dI}{dt}$) must vary smoothly over a region in the image. This provides them with another error, which is based on the spatial variation of $\frac{dI}{dt}$.

Nagel modifies the intensity constraints by including second order intensity variations in the Taylor series expansion. This provides an intensity constraint that is more accurate than the one given above. In addition, at high curvature points along image-contours Nagel's constraint usually provides a unique velocity vector.

2.1.2 Correlation matching schemes

The correlation matching approach begins with the same assumption as the gradient scheme – that the image-intensity of a point remains constant over time, but uses it in an entirely different way. The case that concerns us is the discrete correlation process, since that applies to a discrete sequence of digital images.

Discrete correlation is the process in which an area surrounding a point of interest in one image is "correlated" with areas of similar-shape in a target region in the second image, and the "best-match" area in the target region is discovered. Precise definitions of the terms in quotations are provided below. The center of the best-match-area in the

second image is then regarded as the point corresponding to the point of interest in the first image.

The process of correlation consists of the following steps:

- An area around the point of interest in the first image is chosen as the sample window.
- All the points in a target area (called the *search area*) in the second image, which is expected to contain the match of the point of interest in the first image, are called *candidate match points*. An area identical to the sample window is chosen around each candidate match point. These areas are called *candidate match windows*.
- For each candidate match point a *match-measure* is determined by comparing the image intensities of the points in the sample window and the corresponding points in the candidate match window. The most common match-measures are (i) direct correlation, in which the image intensity values of the corresponding points in the two windows are multiplied and summed, (ii) mean normalized correlation, in which the average intensity of each window is subtracted from the intensity values of each point in that window before multiplication and summing, (iii) variance normalized correlation, which is similar to mean normalized correlation, but in addition the correlation sum is divided by the product of the variances of the intensities in each window, (iv) sum of squared differences, in which the sum of the square of the differences between the intensities at corresponding points is used, and (iv) sum of absolute differences, which is similar to sum of squared differences, but the absolute values of the differences are used instead of their squares.

  In some cases, the match measure may be a weighted sum of the individual point

comparisons. The weights are chosen to increase the contribution of the pixels near the center of the window relative to those of the outlying pixels.

- If either direct, mean normalized, or variance normalized correlation is used to compute the match measure, then the *best match* window is the candidate window that has the maximum value for the match measure. If one of the difference measures is used, then the best match area is the candidate window that minimizes the match measure. The point (in the second image) that is the center of the best match candidate window is regarded as the corresponding-point for the point of interest (in the first image).

At first glance, this technique appears to provide a total constraint on the local displacement vector, i.e., it specifies the displacement vector completely and uniquely. However, this is not always the case. The uniqueness of the displacement vector will depend on the manner in which the match measure varies over the search area. This in turn depends on the underlying structure in the image, viz., whether it is an edge, a uniquely distinguishable structure such as a high curvature point along a contour, or an area of homogeneous intensity. Based on an analysis of the variation of the match measure over the search area, Anandan [Anan84] provides a technique to compute a *confidence measure* associated with each displacement vector.

It should also be noted that the correlation schemes can be fooled in different ways depending on the match measure chosen. For example, if direct correlation is used, then the best match in the search area occurs where the intensity values are high. The difference measures are susceptible to mistakes when the intensity around the point is scaled up or down (see [Hann74,Genn80] for details). The variance normalized correlation is the most

robust measure in the presence of noise and of scale and mean intensity changes. However, it has been noted that certain types of preliminary filtering of the image (e.g., band-pass filtering, see [Burt82]) can provide results of similar quality at a lower computational cost.

2.1.3 A few remarks

It is important at this point to compare the two intensity based schemes for what they compute.

Both schemes fail when the intensity-constancy assumption is incorrect and when the shape of an image area changes due to motion. The correlation scheme is slightly more robust in these situations, since it relies not on an *exact* match of intensities, but on the *best* match over the search area.

Neither scheme performs well when a point gets occluded behind another surface in the image or disappears from the view. The schemes assume that the point is still visible, and so compute an incorrect displacement.

The gradient schemes, since they are based on instantaneous and local image derivatives, are easier to extend to a time sequence of images. The correlation schemes are more cumbersome, since each pair of successive image frames has to be processed separately first, and then their results can be combined.

2.2 Token matching schemes

Token matching schemes for solving the correspondence problem try to avoid the problems that arise when the intensity constancy assumption is violated. This is done by extracting stable symbolic tokens in the images and matching them, rather than depending directly

on the intensity variations.

The tokens can be of varying degrees of complexity according to the structures in the image they represent. The most common are point-tokens, which usually represent some stable and significant image event. The corner point of an occluding contour and the intersection of texture-markings are two examples of such points. The location of the point is usually its primary (and sometimes only) attribute. Other attributes that have been used include the image-intensities in an area around the point (similar to the ideas in correlation based matching), the curvature of the contours at the point, and the location of the point relative to its neighbors.

Sometimes edge tokens may be used, where the location, orientation, and size of the edges are the attributes used to identify and recognize the edges. More complex structural tokens have also been used - for example, the image may be partitioned into regions with bounding contours, and then high curvature points of the contours located. In this situation, the structure is represented as a graph (or a tree), and the graphs (or trees) from the successive frames are matched.

Token matching schemes usually determine the displacement of the token uniquely (in the case of complex tokens, they provide a sort of average motion of the complex structure). However,these local correspondences are error-prone, so a global constraint is also used. Most of the global constraints discussed in section 2.3 will be applicable to any of the token matching schemes.

2.2.1 Point matching schemes

The techniques that use a point-token as a stable matchable feature usually have two parts – extraction and matching.

Point token extraction

The extraction of point tokens can be based on two similar ideas. The first is the notion that points in a highly textured area of the image, i.e., where the intensity variations in multiple image directions are significant enough to produce a structure stable in the face of sensor-noise and area deformation, are useful tokens. The second is the notion that along visible intensity contours in the image (e.g., contours due to albedo changes, or occluding contours between two objects), points of high-curvature of the contours are likely to be stable tokens.

These two ideas are respectively the basis of *interest operators* and *corner detectors.*

Moravec's interest operator Interest operators, as the name implies, attempt to find points in the image that are "interesting". There is clearly no unique definition of this word - in general the definition depends on the algorithm used by a particular operator. The most popular one is known as the "Moravec operator" [Mora80] which works as follows:

- A small area is defined around each point in the image.
- This area is compared to similar areas surrounding all the points within a small radius of this point, excluding itself. Usually the comparison measure is one of the difference measures described in page 15. The minimum of these comparisons is regarded as the interest measure for the point under consideration.

- All points whose interest measure exceeds a certain threshold are candidate interest points.
- Among the candidates a local-maximum selection process is used. This process consists of comparing the interest measure of a point with those of all the other candidate points in a small neighborhood, and retaining it only if it has the maximum interest measure.

Corner detectors Corner detectors attempt to locate the points in the image which correspond to high-curvature (or "corner") points on visible image contours. There are two major ways of achieving this.

1. In the first approach the high-curvature points on the level contours in the image – i.e., the locus of all points with a specific intensity – can be used as the points of interest. Such a curve can be described by an implicit function of the x and y locations of the points that belong to it. High curvature points along such a contour are the locations where the tangent vector of the curve most rapidly changes its direction.

   Kitchen and Rosenfeld [Kitc80] perform an algebraic analysis of such contours, and obtain simple formulas for the *planar curvature* of the level contour at a point in the image. They then proceed to define "corner points" as the locations of the local maxima of the curvature weighted by the magnitude of the intensity gradient vector, i.e., the local maxima of

$$k\sqrt{I_x^2 + I_y^2} = -\frac{(I_{xx}I_y^2 - 2I_{xy}I_xI_y + I_{yy}I_x^2)}{(I_x^2 + I_y^2)}$$

where k is the planar-curvature of the level-contour.

2. Another popular approach for locating corners is to filter the image with a $\nabla^2 G$ operator, locate the zero-crossing contours of the resulting image, and the high-curvature points along such contours. The $\nabla^2 G$ operator can be described as a convolution of the image with a Gaussian mask followed by taking the Laplacian (i.e., $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$) of the convolved image. This process is equivalent to convolving the image with the mask

$$\nabla^2 G(x,y) = c(2 - \frac{x^2+y^2}{\sigma^2}) \exp -\frac{x^2+y^2}{2\sigma^2} \qquad (2)$$

where $c = 1/2\pi\sigma^4$ is a scaling parameter. This convolution is used in many low-level vision algorithms for edge detection and motion analysis. Figure 3 illustrates the mask.

Matching point tokens

Although considerable effort has gone into the careful selection of point tokens, it is surprising that not much has been done to identify stable properties that characterize these tokens. Most algorithms use an area of the intensity-image surrounding the point as its feature, and use one of the correlation techniques described above for matching them. This is surprising, since one of the aims of the token matching process is to use features that remain constant during the movement of the image to find the correspondence between image-points, whereas, as described above, the image intensity values are anything but constant.

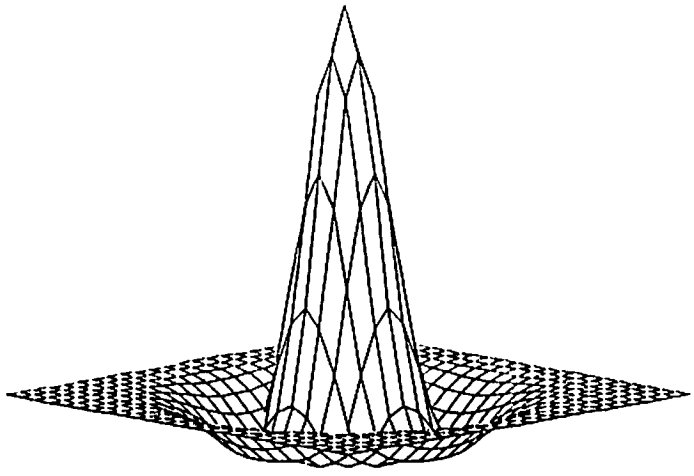

Figure 3: The $\nabla^2 G$ mask. The mask is displayed as a surface plot, in which the height of the surface at any location is proportional to the value of the mask at that location.

Barnard and Thompson algorithm Barnard and Thompson's [Barn80] algorithm is a widely used token matching technique. Initially, each point in the first image (called a "node") is paired with every point in the second image (called a "label") within a preset distance. For each node, associated with each label is the distance between the node and the label (called the "disparity" associated with that label). In addition, each point is also provided a "no-match" label to allow for the possibility that the point does not have a match. Associated with each label is the probability of that match. These probabilities are computed using the variance-normalized correlation measure of two small areas surrounding the two points that are paired. The probability of no match is calculated as the complement of the sum of the probabilities of all the other labels for a point.

These labels provide a partial constraint on the local matches. The global constraint is a consistency condition on the labels of neighboring points in the first image, i.e., the neighboring points must have "similar labels", i.e., labels with nearly equal disparities.

The global constraint is implemented in the form of a relaxation algorithm. This algorithm iteratively updates the probabilities of all the labels for each node. Similar labels of nearby nodes tend to cause an increase in each other's probabilities. The updating process continues (usually for less than 10 iterations), until for each node one of the labels has a significantly higher probability than the others. This label is then considered as the match for that point.

This method of updating probabilities is called *probabilistic relaxation*. It has been shown (see [Humm83]) that this is a type of optimization process that finds the local match labels that are "most consistent" with each other. The measure of consistency of the labels is implicit in the method of updating the probabilities. In this sense, Barnard and Thompson implement a global consistency constraint on the displacement field.

Another matching scheme is described by Prager and Arbib [Prag83]. They extract points from both images and match these points. One important feature is that they allow inexact matches – i.e, the displacement of a point-token in one image is required to bring it *near* a point-token in the other image and not *exactly* to it. A relaxation algorithm is used to compute the displacements which optimize the sum of local match measures and a global consistency measure on the displacements.

2.2.2 Matching edge tokens

Intensity edges or other linear structures in the image can be used as stable features for the correspondence problem. The process consists of two steps – extracting edges, and determining their movement.

In these techniques, an edge is usually specified by its location, orientation, and size.

The *aperture problem* described earlier directly applies to edge-based matching. This is because if an edge is regarded as a small linear structure in the image, there is no local information regarding the amount of movement parallel to the edge. Hence the local matching scheme only provides the movement in the direction normal to the edge (also called *normal flow*).

Most of the edge based matching techniques have been designed for stereopsis. We will describe some of these in section 4. Here we present the technique of Marr and Ullman [Marr81], which is suitable for motion correspondence.

Marr and Ullman's scheme uses the zero-crossings of the $\nabla^2 G$ operator, which we described earlier, as the location of image edges. These are detected by two types of units, one dealing with positive values ("on center"), and the other with negative values ("off center"). On one side of the zero crossing the on-center units (also called S^+) will be active, whereas on the other side the off-center units (called S^-) will be active (see figure 4).

In addition the time-derivative of the $\nabla^2 G$ operator at a point is calculated by T units. T^+ units respond to positive values of the time-derivative and T^- units respond to negative values. The combination of the activity in the different S and T units indicate the direction of movement. For example, in the one-dimensional version shown in figure 4, S^+, T^+, S^- being simultaneously active indicates the presence of a zero-crossing moving from left to right.

This technique provides only the sign of the motion along the direction perpendicular to the edge, i.e, whether it is from right to left, bottom to top, etc. The displacement magnitude – the speed – is not provided. The technique can be slightly modified to provide the speed by comparing the time difference in the activation of neighboring S unit

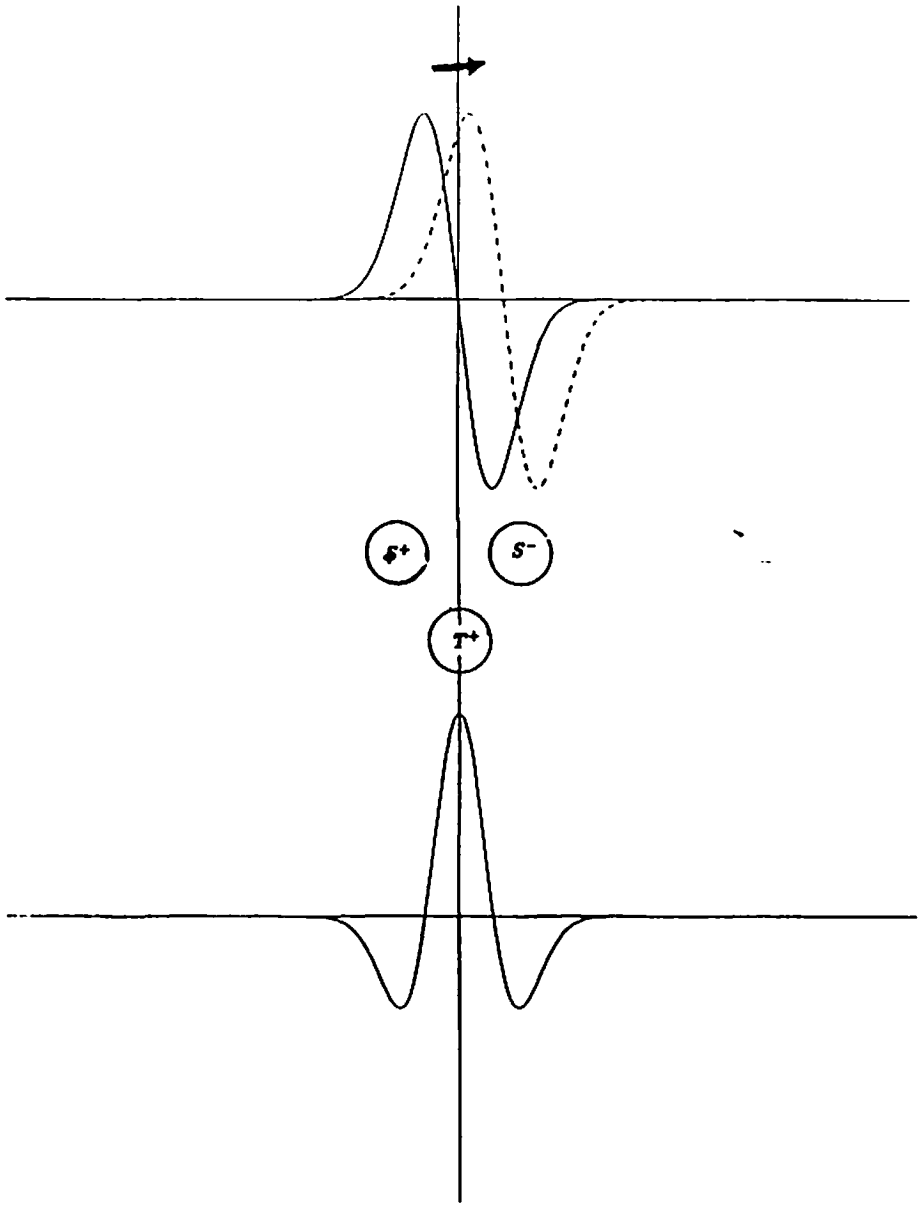

Figure 4: The intensity-profile of the zero-crossing of a moving edge. The top figure shows the result of the $\nabla^2 G$ convolution with a moving step edge, and the bottom figure shows the time derivative of the top figure.

pairs. The time interval between the detection of the zero-crossing at neighboring units located at fixed distances can be used to calculate the speed (see [Marr81] for details).

Just as in the case of gradient techniques, which this scheme resembles in some ways, this technique provides only a partial constraint on the velocity or displacement at a point. A global or semi-global constraint is necessary to compute the complete velocity.

2.3 The use of non-local constraints

Thus far, several different approaches for measuring local correspondence have been described. Some of these provide only a partial constraint on the displacement vector, while others provide a unique but unreliable displacement vector.

As we mentioned earlier, non-local constraints can be used to determine unique reliable displacement vectors. These non-local constraints can be used with almost any of the local correspondence techniques described above. Although some of these have traditionally been intimately used with particular local correspondence techniques, in general it is possible to pair any of the non-local constraints with any of the local correspondence schemes.

The assumption of constant displacement The simplest form of non-local assumption is that the displacement is constant over the image. Such an assumption is strictly true only when the relative motion is a translation parallel to the image plane, and all the environmental points are at the same perpendicular distance from the image plane. However, when restricted to small local neighborhoods and not allowed to completely propagate over the image this assumption is only applied loosely and can be useful.

We illustrate the use of this assumption with the gradient-based local correspondence. It was pointed out that the gradient schemes only provide a partial constraint on the displacement vector, i.e., they constrain the local displacement to a line. The orientation of this line is normal to the local intensity-gradient vector (or parallel to the image edge, if one exists). In a small neighborhood of a point in the image, if the intensity gradient vector changes its orientation, then several intensity constraint lines at different orientations are available. If the displacement is assumed to remain constant in that neighborhood, then the intersection of the constraint lines will be the true displacement. This is equivalent to saying that in a small neighborhood, if there are edges at different orientations (e.g., along an image curve), then their normal velocities can be combined to uniquely determine the velocity of the whole neighborhood.

This idea has been developed by Glazer ([Glaz81]), and Thompson and Barnard ([Thom81]) to compute the displacement field for a pair of images. Their papers also discuss the limitations of this approach.

The assumption of a smooth displacement field The logical step beyond the assumption of constant displacement is to assume the displacement varies smoothly over the image. Later, it will be shown that the true motion of the environmental surfaces can be expressed in terms of six scalar parameters and the distance of each point (or its "depth") from the image plane. If the depth of the environmental points is assumed to vary smoothly across the image plane, the displacement vector field must also vary smoothly. This is the basis for any of the smoothness assumptions.

The most common form of the smoothness assumption is the minimization of a smooth-

ness error which measures the spatial variation of the displacement field. Such a measure usually includes the partial derivatives of the displacement field. The measure can be over an area of the image [Horn80,Anan85], or only along contours in the image [Hild83,Nage83b]. The latter case is used in an attempt to eliminate the propagation of the smoothness constraint across depth or object boundaries in the image. The image contours usually trace such boundaries and restricting the smoothness constraint to be parallel to them and not across them may have the desired effect.

An example of an area-based smoothness error is

$$E_{smoothness} = \int\int (u_x^2 + u_y^2 + v_x^2 + v_y^2) dx dy$$

This form is due to Horn and Schunck [Horn80]. Horn and Schunck also formulated an approximation error

$$E_{approx} = \int\int (I_x u + I_y v + I_t)^2 dx dy$$

which measures the deviation of the local displacement from the intensity constraint line.

Horn and Schunck, and many others following them, attempt to minimise a sum of the two errors

$$\alpha^2 E_{smoothness} + E_{approx},$$

where α is a weighting factor, to obtain the displacement vector field. The minimization process is usually in the form of a relaxation algorithm which iteratively modifies each displacement vector according to the values of its neighbors and the local approximation error.

Anandan and Weiss [Anan85] provide a modified form of the approximation error, based on a correlation matching algorithm. They express the local displacement vector U in

terms of a local basis (e_{max}, e_{min}). The initial displacement vector approximation provided by the matching algorithm is D. The quantities c_{max} and c_{min} are confidence measures associated with the components of D along the directions e_{max} and e_{min} respectively. These confidence measures are also provided by the matching process. The approximation error used by Anandan and Weiss is

$$E_{approx} = \sum c_{max}((U - D) \cdot e_{max})^2 + c_{min}((U - D) \cdot e_{min})^2$$

which can be regarded as generalizing of Horn and Schunck's scheme so as to recognize unique displacement vectors wherever available (e.g., intensity corners).

Finally, an example of a smoothness constraint that is applied along a contour is seen in the formulation of Hildreth [Hild83]. Given the normal velocity along a contour (due to the intensity-constraint or by edge matching schemes), she minimizes

$$E = \int (|\frac{\partial U}{\partial s}|^2 + \beta(U \cdot e_n - D \cdot e_n)^2) ds$$

where U is the desired velocity vector, e_n is the unit vector normal to the contour, $(D \cdot e_n)$ is the velocity component normal to the image contour, and s is the arc-length along the contour. In this way, she minimizes the variation of the velocity along the contour while also minimizing the deviation of the normal component from its prior estimated values.

The smoothness constraints described here have the advantage of being in a rigorous mathematical setting and so can utilize some known methods of solving optimization problems. Unfortunately, however, none of these heuristics on the variation of the displacement fields are likely to be precisely true. Indeed, they are not even based on the geometric transformations that are physically possible during motion. It is conceivable that

a more carefully formulated heuristic would lead to the computation of a more accurate displacement field.

The smoothness assumption is invalid both at object boundaries, because the different objects may move independently of each other and hence not have the same motion parameters, and as well as at depth discontinuities in the environment, since the discontinuities in depth cause corresponding discontinuities in displacements even if the motion is the same. Applying smoothness constraints across such boundaries has severely detrimental effects, since the displacement fields on either side of a boundary should not directly influence each other. A number of researchers point this out, and suggest prior detection of the location of such discontinuities as a way of solving the problem. However, no technique has yet been able to achieve this. This is indeed a serious limitation on the use of these smoothness constraints, and one that will be a focus of the research in this area.

2.4 Spatial frequency analysis

The approaches described above for solving the correspondence problem usually work only when the displacement is small or (in the case of techniques measuring image-velocities) if the time difference between the two frames is small. If the displacements are large, intuitively it would seem to be useful to track or match large structures in the image, since these would be uniquely identifiable over a distance. The following situations help to explain the key ideas of this section.

1. Consider a highly-textured region of an image, the texture being fairly regular and having a small period (i.e., the "spatial-frequency" of the intensity in that region is high). Assume that the region is displaced by a large amount (much greater than the

period of the texture). In this case, if we focus our attention on a small area in the middle of the region, there is no way to accurately measure the displacement of this area. This is because the texture is repetitive and we can only detect motion modulo the period of the texture. However, if a rough estimate of the displacement is known (with an error less than the period of the texture), the high-frequency information can be used to obtain more precise estimates.

2. Consider another region where the texture has a larger period and the image intensities vary slowly over the region. This implies that the region has no sharp edges that can be clearly identified and localized. Although an estimate of the movement of this region can be obtained, the inability to localize image-events implies that the estimate will not be very precise. This problem is even worse in the presence of noise.

From a computational viewpoint, these observations suggest that the image should be decomposed into its spatial frequency components. The low frequency components can be used to obtain rough displacement estimates (over a large range of possible displacements) and the higher frequency components can then be used to localize these estimates.

This idea is familiar to psychologists [Adel83]. In computer measurement of displacements of image points, it appears in early stereopsis formulations by Marr and Poggio [Marr79]. A detailed description can be found in [Burt83,Glas83].

An attempt to formulate an efficient computational technique based on this idea should also take into account the scale and resolution of the image information. As mentioned above, when displacements are large, we must rely on low-frequency image information for their measurement. This measurement cannot be very precise, since any of the measures

used on the low-frequency information will not be sensitive to small variations in the displacement. In addition, the need to measure large displacements implies that a large area of the image must be searched. Taken together, these suggest that low-frequency information should be represented at coarse spatial resolutions, and that the large displacements should be measured using a large scale.

In a similar manner, high frequency information should be represented at a fine resolution and used for measuring small displacements at a small scale.

The techniques that use these observations usually pre-process the image using a set of spatial frequency band-pass filters, each an octave wide and an octave apart from each other. The filters are usually achieved through convolutions with a family of $\nabla^2 G$ masks (as in equation 2), with increasing σ values corresponding to decreasing center frequencies. These filters are also called channels, and the output of each of these filters are represented at a resolution corresponding to their Nyquist sampling-rate. With the octave-wide channels, this results in a set of images whose resolutions successively increase by a factor of 2.

These ideas are pursued in detail by [Burt83], [Nish84], [Glaz83], [Anan84], [Quam84], who use them in various techniques for solving the correspondence problem. The basic approach involves applying one of the correspondence techniques described earlier in this section on each of the spatial frequency channels. The details of the communication between the channels is usually depends on the technique and will not be discussed here.

3 DETERMINING 3-D MOTION AND STRUCTURE

The primary goal of motion analysis is to determine the 3-dimensional structure of the objects in the environment and the relative movement of the camera and the objects in the scene. The determination of the 2-dimensional image displacements or velocities of the image-points is only one (although an important one) of the steps involved. The interpretation of the displacement (or velocity) fields to determine the 3-d structure of the environment and the relative 3-d motion between the objects and the camera is another important step. As mentioned before, it may even be possible to directly determine the 3-d structure and motion without computing correspondence of points or other local image events.

We begin by noting that the instantaneous movement of any rigid object can be described as the combination of a rotation and a translation with reference to any given coordinate system. The rotation is usually expressed as an angular velocity ω about an axis oriented along the unit vector $\vec{e}_\Omega$ and the translation as a 3-d vector $\vec{T}$. It is also common to represent the rotation as a single vector $\vec{\Omega}$ of length ω and direction that of $\vec{e}_\Omega$.

The choice of the coordinate system is arbitrary, since the rotation and translation vectors with respect to two different coordinate systems are related by a simple geometric transformation.

The 3-d structure of the visible environment is completely specified if we know the distance along the optical axis (the "depth") of each point in the image. This, however, may not be the most useful form. If the task at hand is to describe the 3-d shapes in terms of known geometrical objects (such as cylinders, spheres, etc.), there must also be

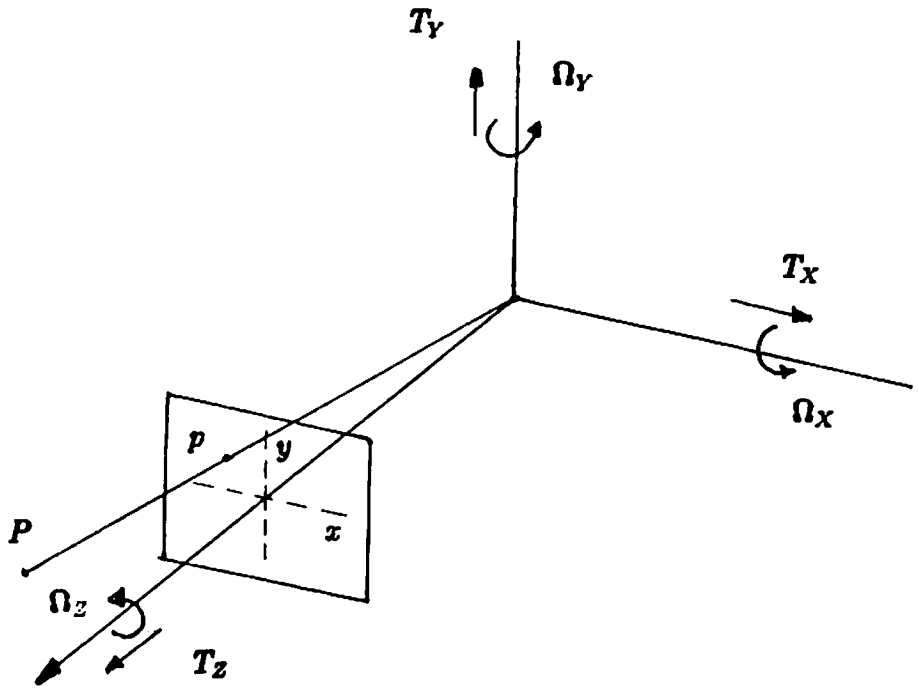

Figure 5: The camera coordinate system

a combining step that transforms these pointwise descriptions to that of solid shapes. However, in this section we restrict our attention to pointwise 3-d location information, since that has been the approach taken by most researchers.

3.1 The geometry of flow fields

The geometrical nature of the optical flow fields can be understood through a series of equations that relate the coordinates of the image-points and the motion parameters to their velocities.

Let (X, Y, Z) be a cartesian coordinate system affixed to the camera (see figure 5) and let (x, y) represent the corresponding coordinate system on the image plane. The focal length of the camera is assumed to be known, and can be normalized to 1, without loss of generality.

Consider a point P on the object, located at $\vec{P} = (X, Y, Z)$. The 3-d velocity $\vec{V} =$

$(\dot{X}, \dot{Y}, \dot{Z})$ of the point is given by

$$\vec{V} = \vec{\Omega} \times \vec{P} + \vec{T} \tag{3}$$

where $\vec{\Omega} = (\Omega_X, \Omega_Y, \Omega_Z)$ is the rotation vector and $\vec{T} = (T_X, T_Y, T_Z)$ is the translation vector, whose direction and magnitude specify the direction of translation and the speed respectively.

The task of determining the 3-d motion of an object can be described as the task of recovering the parameters $\vec{\Omega}$ and $\vec{T}$.

If $\vec{p} = (x, y)$ is the image position of the projection of P and $\vec{U} = (u, v) = (\dot{x}, \dot{y})$ is the image-velocity of that projection, then using

$$\begin{aligned} x &= X/Z \\ y &= Y/Z \end{aligned} \tag{4}$$

we find from equation (3)

$$\begin{aligned} u &= -\Omega_X xy + \Omega_Y(1 + x^2) - \Omega_Z y + (T_X - T_Z x)/Z \\ v &= -\Omega_X(1 + x^2) + \Omega_Y xy + \Omega_Z x + (T_Y - T_Z y)/Z \end{aligned} \tag{5}$$

We will refer to these equations as the *optic flow equations.* They [1] apply only to the velocities of the image points and not to the displacements of the points in a discrete image sequence. However, when the field of view, the amount of rotation, and the translation in depth (i.e., T_Z in the above equations) are all "small" the image displacements are good approximations to the instantaneous velocities. Although some approaches deal explicitly

[1]Note that these equations are based on the choice of the camera-based coordinate system as the frame of reference. This is not a restriction, since the choice of the frame of reference is arbitrary, and does not change the interpretation of the flow field. The values of the parameters of motion and structure depends on the reference frame, but given two frames of reference the transformation of these values between them is a fixed.

with displacements most of the techniques for determining 3-d structure and motion of the environment use these approximations.

Understanding the equations

Six parameters describe the motion of an object and three parameters describe its 3-d structure. The three components each of $\vec{T}$ and $\vec{\Omega}$ specify the relative motion of the object and the camera. The X, Y, Z coordinates of all the points on the object together specify the structure of the object.

The known image-position (x, y) of a point on the object specifies the direction of $\vec{P}$. Hence, only the distance $|\vec{P}|$ of P along that direction is unknown.

In the optic flow equations, the components of T (T_*) always appear in the form $\frac{T_*}{Z}$. This means that based purely on the image velocities, we cannot determine the absolute translational velocity and the absolute distance of a point along the line of sight. They can both be multiplied by the same scale factor k without changing the optic flow field. Intuitively, this means that given an image and its optic flow field, if all the objects in the world are moved away from the camera by a factor k, the object magnified k and the relative translational velocity is multiplied by k, the resulting flow field will be identical to the original field.

The velocity $\vec{U}$ of an image-point can be expressed as $\vec{U} = \vec{U}_R + \vec{U}_T$ - the sum of its rotational and translational components. From the optic flow equations, it can be shown that the rotational component is not influenced by Z, whereas the translational component is. This suggests that the rotational component of the optic flow field will not be useful

in determining the 3-dimensional structure of an object. The translational component contains all the available information regarding the structure.

As explained before, the parameters of motion typically do not vary from point to point in the image. All the points on a rigid object undergo the same motion and have the same motion parameters. [2] Hence the number of parameters of motion are few, one set corresponding to each area of the image having an independent relative motion with respect to the camera. When only the camera moves, the whole image forms one coherently moving area.

On the other hand, unless some assumptions are made regarding the structure of the environment, there is one unknown Z value for each image-point. Many techniques often assume that the environmental surfaces can be approximated by piecewise planar or quadric surfaces in order to simplify the computation of structure.

There are three major approaches that are of interest to us. The first type does not require prior computation of optic flow – in fact the optic flow can sometimes be obtained simultaneously with the 3-d motion parameters. Often, these techniques apply only to restricted camera motion (or a restricted motion of the scene as a whole.), and do not allow independently moving objects. The second type of technique requires knowing the correspondences for a few points in the image. These also usually do not allow independent object motions. The third type of technique requires an optic flow field. One such technique allows multiple independently moving objects.

[2] If the object is non-rigid the situation is more complex. Most of the current work applies only to rigid motion. Hence the same restriction will apply here.

3.2 Processing a restricted class of motions

The problem of processing a restricted class of motion to obtain directly the parameters of motion and image structure without having a prior solution to the correspondence problem has been dealt with extensively by Lawton [Lawt84]. All of the cases Lawton considers are situations where the motion is solely due to that of the camera. The class of motion processed includes pure translation of the camera in an arbitrary direction, pure rotation of the camera about an arbitrary axis passing through the focal point, and known-planar motion – one in which all the environmental displacements are restricted to lie on the same plane. This last case arises when the axis of translation $\vec{T}$ lies on the plane perpendicular to the axis of rotation $\vec{\Omega}$.

These cases considered by Lawton are significant because many practical situations with a moving camera fall into one of these cases - e.g., a pilot attempting to land usually follows pure translational motion, while a car moving and turning on a road is a case of planar motion where the motion is on the ground plane.

In each of these cases the number of unknown parameters is small and hence computation can proceed easily. The assumption that all the observed motion is due to the movement of the camera allows us to treat the whole image as a single rigid object. This enables information from everywhere in the image to be used for the recovery of motion parameters, an idea that leads to a robust technique.

The next two sections describe briefly the approach used by Lawton for the case of pure translation and the case of pure rotation. The case of known-planar motion is similar, and is not included here.

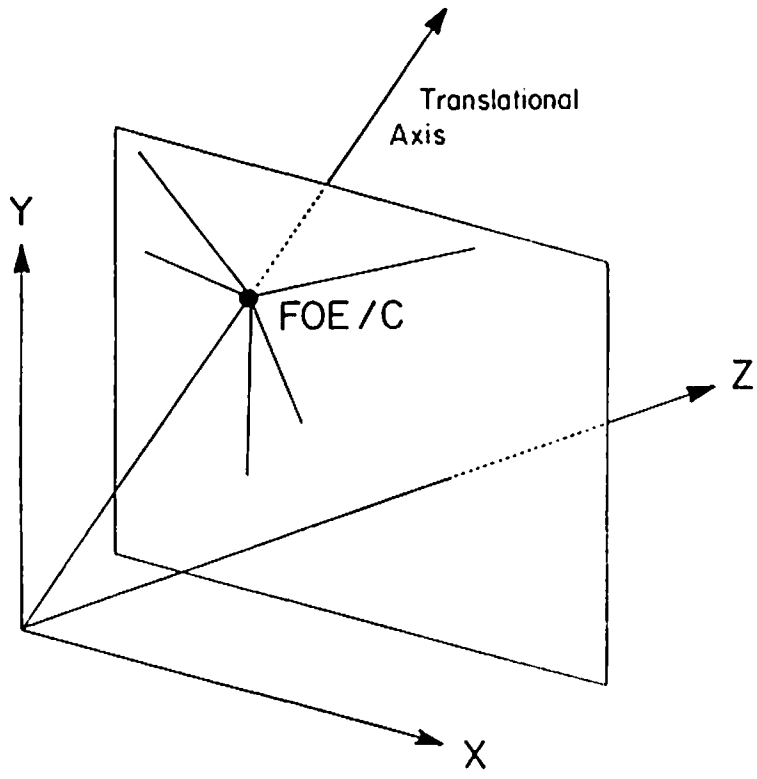

Figure 6: The displacement due to camera translation

Pure translation

When the camera motion is a pure translation towards the environment, all the displacements on the image appear to emanate radially from a single point in the image. This point is known as the **Focus of Expansion** (FOE). It is also the point of intersection of the axis of translation with the image plane. This is illustrated in figure 6. If the camera moves away from the environment, the displacements appear to converge at a point on the image plane called the **Focus of Contraction** (FOC).

In the case of pure translation, the problem of determining the motion of the camera reduces to that of locating the FOE or, equivalently, the axis of translation. In either case, the number of parameters is two, thus greatly simplifying the problem of general motion, which has six parameters. Additionally, knowing that all the displacements have to lie along the radial lines from the FOE provides a powerful constraint that simplifies the correspondence problem.

The displacement ΔD of the image of the projection of a point in the 3-d environment is directly proportional to the distance D of the projection from the FOE and inversely proportional to the distance Z of the point from the camera. More precisely,

$$\frac{\Delta D}{D} = \frac{\Delta Z}{Z} \tag{6}$$

$D, \Delta D$ and Z are as defined above, and ΔZ is the displacement of the camera along its optical axis.

If the FOE is known, then D is known and ΔD can be measured. From the equation above, it is clear that only the ration Z and ΔZ can be recovered. It is common practice to set ΔZ to 1 and then obtain the depths. Alternately, some point in the image can be arbitrarily chosen to be at unit depth and then the relative depth of the others can be obtained.

Based on these observations, Lawton provides a simple algorithm for the location of the FOE and the computation of relative depth. Instead of searching for the FOE, Lawton searches for the direction of translation. This way the search is conducted in a unit sphere surrounding the focal point. Each point on the surface of the sphere corresponds to a direction of translation.

Given a hypothesized direction of translation, the corresponding FOE can be determined. Given a set of points S in one image (which are chosen by an interest operator similar to those discussed in section 2.2.1), each point is matched with points in the other image which lie along the radial line from the FOE. Associated with each potential match of an "interesting" point is an error measure. Lawton's error measures are based on the correlation measures described in section 2.1.2. Of all the candidate-points along the radial

line from the hypothesized FOE within some fixed maximum displacement, the one that minimizes this error measure is chosen. Let $e(i, \vec{T})$ be the minimum error-measure for the point i in the set S under the hypothesized axis of translation $\vec{T}$. Then, Lawton defines the error measure $E(\vec{T})$ as

$$E(\vec{T}) = \sum_{i \in S} e(i, \vec{T})$$

The correct axis of translation is the one that minimizes E. The details of the search for the minimum can be found in [Lawt84].

Once the true axis of translation and the corresponding FOE are determined, then measurement of the displacement along the radial line immediately provides the relative depth of each point under consideration.

Pure rotation

When the motion of the camera is a pure rotation about an arbitrary axis, each image point follows a path that is a conic. The exact curve along which the point travels is the intersection of the image plane with a cone passing through the image point, whose vertex is at the focal point, and whose axis is the same as the axis of rotation. A typical case of rotation is illustrated in figure 7

Given a hypothesized axis of rotation, the path of each point can be determined. In addition, the angular displacement of the point along this path is the same for all image points, regardless of their depth. These facts are used by Lawton in an algorithm that searches for the axis of rotation. An error measure similar to that of the translational case is defined. The additional constraint that the angular displacements must be identical is

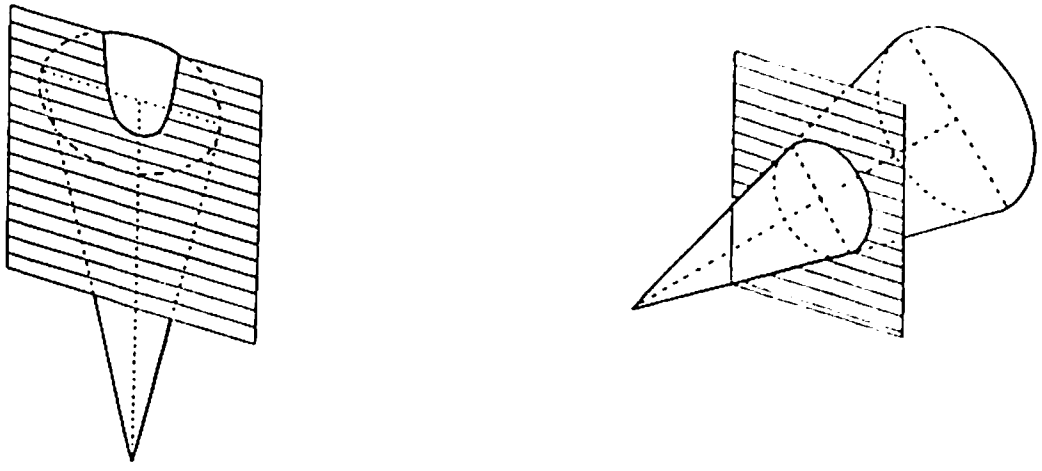

Figure 7: The path of an image point due to camera rotation. The two figures display the cases when the axis of rotation is parallel to (the figure on the left) and perpendicular to (the figure on the right) the image plane

also used to improve the search.

3.3 Approaches based on point correspondences

A class of techniques exist that uses the knowledge of the correspondence of a few points in the images to determine the 3-d structure and motion of the objects in the scene.

The fundamental assumption that is used (although often implicitly) by all such techniques was noted by Ullman. This is the **rigidity assumption**, which states that "any set of points undergoing a 2-d transformation between images which has a unique interpretation as a rigid object moving in space should be interpreted as such." [Ullm79]

As such, it is clear that displacement of each image point is a function of the motion parameters (six in number) and the depth of the point. It is then a matter of obtaining a sufficient number of points and their displacements to be able to solve for these unknown parameters and depths. This is the approach used by all techniques that rely on known

point correspondences. A few prominent algorithms that use this approach are outlined briefly here, leaving aside details of the algorithms.

It should be noted that all these techniques assume that the points involved in the computations belong to the same object. Ullman describes a possible way of avoiding this assumption, but it is not clear if his approach works successfully in practice.

Ullman's approach

Thus far, all the discussion in this chapter has concerned the analysis of a dynamic image-sequence generated under perspective or central projection. However, Ullman separately considers both orthographic and perspective projection images [Ullm79]

For orthographic projection, Ullman derives the following theorem, which he calls the **structure from motion** theorem:

Given three distinct orthographic views of four points in a rigid configuration (and the correspondences between their image locations), the structure and motion compatible with the three views are uniquely determined

For perspective projection Ullman derives the condition that three views of five points are usually sufficient. However, he notes that greater accuracy is needed in their image locations (in the perspective case) to achieve accurate results, and the computation required to compute the structure and motion is more complex.

Ullman also claims that the results in the perspective case are superior to human performance. He suggests an interpretation based on a polar-parallel-projection assumption. Polar-parallel-projection assumes that in a small area of the image one can assume orthographic projection if the objects imaged in that area are sufficiently far away from the

camera. This approach is claimed to produce results comparable to human performance.

The approach of Roach and Aggarwal

Roach and Aggarwal [Roac80] base their analysis on the equations of perspective. With reference to some arbitrarily chosen cartesian coordinate system, the three dimensional world coordinates of a point can be expressed as

$$\begin{aligned} x &= X_0 + \frac{F}{(F - z')}(a_{11}x' + a_{12}y' + a_{13}F) \\ y &= Y_0 + \frac{F}{(F - z')}(a_{21}x' + a_{22}y' + a_{23}F) \\ z &= Z_0 + \frac{F}{(F - z')}(a_{31}x' + a_{32}y' + a_{33}F) \end{aligned} \tag{7}$$

where (X_0, Y_0, Z_0) are the coordinates of the camera location, $(a_{11}, \cdots, a_{33})$ are functions of the three orientation parameters of the camera, (x, y, z) are the 3-d coordinates of the point, (x', y') are the coordinates of the image of the point on the focal plane, F is the focal-length of the camera, and z' is a free-variable.

Equations (7) give the locus of points that form a straight line in space passing through the camera origin (X_0, Y_0, Z_0) and the image of the point at (x', y') on the focal plane. The location of the point along the line is determined by the free parameter z', which can be arbitrarily specified. Specifying z' has the same effect as changing the scale of the global coordinate system – similar to choosing the scale-factor k discussed in section 3.1.

Roach and Aggarwal also determine the projection equations giving the image coordinates of the point as functions of its world coordinates and the camera parameters. These are the standard perspective projection equations, and are omitted here.

Although it appears that in this method, there are six unknown parameters for each

of the two camera positions, the choice of the global coordinate system is arbitrary. We can choose the global coordinate system to coincide with the camera coordinate system (described in figure 5), of one of the camera positions, thus reducing the number of unknown camera parameters to six. There are also three unknown coordinates for each point. The the scale factor z' can be fixed by arbitrarily choosing the z location of any one point.

Roach and Aggarwal show that by choosing five points in one image whose corresponding locations in the other image are known, we can obtain 18 non-linear equations with 18 unknown parameters. These 18 parameters include the camera parameters as well as the 3-d coordinates of the points. The equations are solved using an iterative technique.

In general this method is severely affected by noise and errors in the correspondence process. They claim that they need two views of 12 points in order to give robust measurements of structure and motion.

The approach of Tsai and Huang

If (x, y) and (x', y') are the coordinates of the projection of a point in two images, Tsai and Huang [Tsai84] derive the equation

$$(x \quad y \quad 1)\, E \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = 0 \tag{8}$$

where E is a 3×3 matrix with 8 unknown entries and an unknown scale factor. The entries of E are linear functions of the motion parameters. The structure parameters are not involved in E.

Given n points whose correspondences are known, Tsai and Huang present a technique that uses a least-square error method to determine E, followed by a singular value de-

composition of E to determine the motion parameters. Once the motion parameters are known, the determination of the structure is straight-forward.

This method is unusual in that it is one of the very few attempts to solve the problem using a set of linear equations. This is claimed to improve the stability and noise sensitivity of the results.

3.4 Techniques requiring optic flow

If the velocities (or displacements) of a fairly dense set of points in the image are given, there are a number of methods for computing the 3-d structure and motion. Some of these methods require one flow vector for each point (or pixel) in the image, whereas others require that the available flow vectors be spread over the image without requiring they be known everywhere. Because all these techniques require more than a small number of point correspondences, they have been grouped here as those requiring an optic flow field.

There are two ways in which the optic flow fields can be used in this process. The local derivatives (or spatial differences) of the flow vectors can be used to provide information about the structure and motion of the object. Alternatively, some global measures of the flow vectors can be used.

Local techniques have the advantage that their computations are based on local properties, and so the distant areas of the image (which are often parts of different objects) can be treated independently. The disadvantage of the local techniques is that they usually require local computation of up to second order derivatives of the flow vectors. This is usually difficult or impossible to obtain using currently available methods, and the process of differentiation is highly sensitive to even small inaccuracies. In the case of global

techniques, the situation is just the reverse – while local errors do not severely affect the process, it requires global computations which lead to communication bottlenecks and slow processing.

3.4.1 Local techniques

As the camera moves relative to a surface patch in the environment, the image of the patch moves and deforms in shape. The local deformation and movement of a small area can be used to determine the parameters of motion and the surface structure. This is the fundamental observation used by practically all local techniques.

Two approaches are outlined here to introduce the reader to the relevant work in this area. One type of approach is primarily due to Waxman and Ullman [Waxm83,Waxm84a,Waxm84b] and the other due to Longuet-Higgins and Prazdny [Long80] and Rieger and Lawton [Rieg83]. The approaches of Longuet-Higgins and Rieger are directly related to some of the mathematical analysis of Koenderink and Van-Doorn on optic flow fields generated by a moving observer and a stationary environment. This last work here is complicated, and primarily of theoretical interest. Hence, it will not be described here.

Waxman's approach The approach of Waxman and his colleagues is best summarized as what he calls an **image flow paradigm**. Consider an environmental surface-patch (planar or quadric) moving rigidly through space relative to an observer. Local flow of the images of the points on the surface can be described in terms of 12 **flow deformation parameters** evaluated at the image of any point on the surface. The origin of the reference coordinate system is chosen at that point. The twelve quantities consist of the two

components of the image velocity of the point, the three independent strain rates, the spin, and the six independent derivatives of the strain rate and the spin. If the two components of the optic flow field are expressible as Taylor series expansions around the point, then these parameters are linear combinations of the first six coefficients of the two series.

The strain-rate consists of the rate of stretch of image lines oriented along the x and y axes, and the rate of change of the angle between two line segments oriented along the axes. The spin is the rate of rotation of a small neighborhood of a point in the image around that point.

These 12 parameters can be used to compute locally the six parameters of motion and the structure parameters of the surface. The general case of a quadric surface has six parameters, corresponding to the six coefficients of the Taylor series expansion of the depth around that point. Physically, these are the location, the two slopes, and the three curvatures of the surface-patch. All of these are shown to be specified only up to a scale factor.

The deformation parameters can be obtained by differentiating a dense flow field around the point of interest. This is likely to be error-prone, since it relies on obtaining reliable flow fields – as yet an unsolved problem.

Alternatively, Waxman and colleagues propose an **evolving contour analysis** which attempts to measure these parameters directly from the image. This approach involves detecting contours on the image that correspond to stable physical markings on the surfaces and studying the manner in which these contours deform over a sequence of image frames. It should be noted here that there is no known technique to do so.

The above ideas are relevant for a single surface patch in rigid motion relative to the

camera. When dealing with multiple moving surfaces (both when the surfaces are from different parts of the same object, or from different objects), Waxman's approach consists of separating the areas of the image within which the flow field is twice differentiable. Once again, as noted in the previous section on the correspondence problem, this is one of the major unsolved problems of motion analysis.

Thus, while Waxman's approach is impractical at present, the ideas are useful in two ways. First they are theoretically significant in that they analyze local properties of the image flow and relate them to the motion and structure of physical surfaces. Second, their analysis provides motivation for the direct measurement of image deformation parameters, an approach that is different from the popular paradigm of making displacement or velocity fields as the sole basis of further computation. The difficulties in determining accurate displacement fields makes this an attractive alternative.

Using motion parallax In stereopsis, parallax is the disparity between two points at different depths but at nearby visual directions. As noted earlier, the magnitude of the displacement of a point due to the translation of an object relative to the camera is inversely proportional to the depth of the point. Further, it was also noted that the displacement takes place along the line joining the FOE and the point. Thus, two different points at different depths but nearby visual directions will be displaced in almost the same direction but by different amounts. This is known as motion parallax.

If now there is also a rotational component to the motion, the displacement due to this component is independent of the depth of the points. Therefore when the difference between the displacements of nearby points is considered, their rotational components

cancel each other. Hence, the difference between the displacement vectors of the two points is due only to the translational components and will point toward or away from the FOE.

This observation is central to the method of Longuet-Higgins and Prazdny [Long80] and of Rieger and Lawton [Rieg83]. Longuet-Higgins' formulation is for the ideal situation where the two points are infinitesimally apart in the image. Rieger and Lawton, on the other hand consider the situation when the displacement vectors may be those of points that are a finite distance apart. Their analysis takes into consideration the inaccuracies that may be introduced due to the distance between the points.

If many such point pairs are chosen the intersection of the difference vectors will indeed be the FOE. This idea is used by Rieger and Lawton. Point pairs are chosen from the displacement vector field such that the two points are separated by a small distance and have reliable displacement vectors. (They use an algorithm described in [Anan84] to obtain the displacement vectors and reliability measures.) But among the difference vectors, only those above a certain threshold are maintained. The best intersection of these difference vectors is then determined as the FOE.

Once the FOE is known the axis of translation is known. The rotation can be then be computed simply by removing the translational component from each displacement vector and searching for a rotation $\vec{\Omega}$ that gives the appropriate rotational components of the displacements.

This method implicitly assumes a stationary environment and a moving camera. It is not clear it can be generalized to situations involving multiple moving objects.

3.4.2 Global techniques

Whereas local techniques rely on the local differential properties of optic flow to provide information about the environment, global techniques recognize the fact that the motion parameters are the same for an entire rigid object, and attempt to recover them.

If the optic flow information from distinct parts of the object can be brought together in a coherent way, it can be used to identify the six parameters of motion that simultaneously give all the flow vectors. There are several attempts to do this, but the one that is of greatest interest is the technique of Adiv [Adiv85]. His is one of the very few attempts to deal with images of scenes containing multiple independently moving objects.

Adiv takes a two stage approach. The first stage consists of grouping local flow vectors into those consistent with the motion of a planar patch. In the case of an arbitrary planar patch, it can be shown that the flow field is a quadratic function of the image coordinates (x, y):

$$\begin{aligned} u(x,y) &= a_1 + a_2 x + a_3 y + (a_7 x + a_8 y)x \\ v(x,y) &= a_4 + a_5 x + a_6 y + (a_7 x + a_8 y)y, \end{aligned} \tag{9}$$

where the $(a_1, \cdots, a_8)$ are functions of the slopes and the location of the planar patch, as well as of the six motion parameters. These equations represent what Adiv calls a Ψ transformation – a mapping of the two dimensional image onto itself.

Adiv notes that an environmental surface can be approximated piecewise by planar surfaces, provided that the distances between the real surface and the approximating planes are small compared to the distances of these surfaces from the camera. In this case, the flow vectors are grouped into those consistent with the rigid motion of a planar patch.

The grouping process itself consists of two parts. If the second order terms are ignored

then the flow vectors are consistent with an affine transformation of the image patch. These affine transformations are parameterized by the 6 quantities $(a_1, \cdots, a_6)$ and, further, each component of the flow vector is only a function of 3 parameters. This observation enables Adiv to use a generalized Hough transform to determine the 6 parameters consistent with flow vectors in a patch. [3]

In the second part of the grouping process adjacent segments consistent with the same Ψ transformation are merged together as planar patches in rigid motion. Thus the grouping process also enables the separation of distinctly moving objects (or surfaces) into different "segments" in the image.

In the second stage of Adiv's process, segments whose optic flow vectors are all consistent with a single rigid motion are grouped together as a single object. The motion and the structure parameters of a group of segments is determined as those that minimize an error $E(\vec{\Omega}, \vec{T}, \{Z_i\})$, where $\{Z_i\}$ are the depth values of all the points in the group. For any $(\vec{\Omega}, \vec{T}, \{Z_i\})$ the error E is defined as the weighted sum of squares of the differences between the flow vectors predicted by these parameters and the given flow vectors of the points in the group. The weights are confidence measures associated with the individual flow vectors.

Since the depth values and the magnitude of the translation vector can be specified only up to a scale parameter, the task then is one of finding the direction of translation $\vec{e}_T$, the relative depths $\{\tilde{Z}_i = (Z_i\ |\vec{T}|)\}$, and the rotation parameters $\vec{\Omega}$ that minimize E.

[3] The Hough transform is a global parameter searching process. It is a voting process, where each piece of data (in this case the flow vector) votes for all the parameters that are consistent with it. Only the parameters that are consistent with a large part of the data will get significant votes. Local maxima of votes in the parameter space indicate possible true parameters and their contributors in the data usually correspond to a consistent set. This is a well known technique in pattern recognition and computer vision (see [Duda73,Ball82]).

The minimization process consists of the following steps: Adiv first shows that the optimum set of $\{\tilde{Z}_i\}$ values can be written as a function of the motion parameters. This allows him to recompute a new error $\sigma(\vec{e}_T, \vec{\Omega})$ as a function solely of the motion parameters. For each hypothesized direction of translation, the rotation parameters that minimize σ are determined. Substituting these into σ yields a new error function σ' which is a function only of the direction of translation:

$$\sigma'(\vec{e}_T) = \min_{\vec{\Omega}} \sigma(\vec{e}_T, \vec{\Omega})$$

The direction of translation that minimizes σ' is then chosen as the optimal estimates of the direction of translation, and the corresponding rotation parameters as the optimal estimates of the rotation parameters. The relative depths are then computed in a straightforward manner.

A final stage of verifying the hypothesized groupings of the segments is also incorporated. The details on the minimization algorithm and the hypothesis verification phase can be found in [Adiv85].

3.5 Concluding remarks

The problem of finding 3-d structure from motion appears to be as difficult as the problem of measuring the point correspondences. There appear to be robust techniques that apply to cases of restricted sensor motion, but the general problem of dealing with multiple moving objects is still difficult to solve. Most of the techniques deal with the information from two frames. Usually, they try to provide the depth of each point in order to describe the structure, and the 3-d motion parameters in order to describe the motion.

It would appear that in the longer run, a shift of emphasis towards a qualitative description of the motion is perhaps more useful. In addition, the information from a sequence of images should be used, and a gradual refinement of the structure of the environment over time is likely to prove useful. These are open issues for research in this domain.

4 STEREOPSIS

The analysis of a pair of stereo images is in many ways simpler than motion analysis. For instance, the knowledge of the relative locations and orientations of the cameras has been used to reduce the efforts involved in finding correspondences of image-events. The interpretation of the disparity information is also simplified, since the only unknown parameters are the depths of the image-points.

Stereopsis is also perhaps one of the best understood aspects of human vision. In fact, there are a number of techniques which claim to be computational models of human stereopsis. The results from these techniques are comparable to results from psychophysical studies of human vision.

In this section, we first overview the geometric issues involved in stereopsis and describe how the matching-process can be simplified. We then describe two techniques that utilize general physical and geometric constraints, and two systems that use information specific to a scene-domain. Following this, we review the major attempts at modelling human stereopsis. Finally, we describe the issues involved in the 3-d interpretation of disparity data.

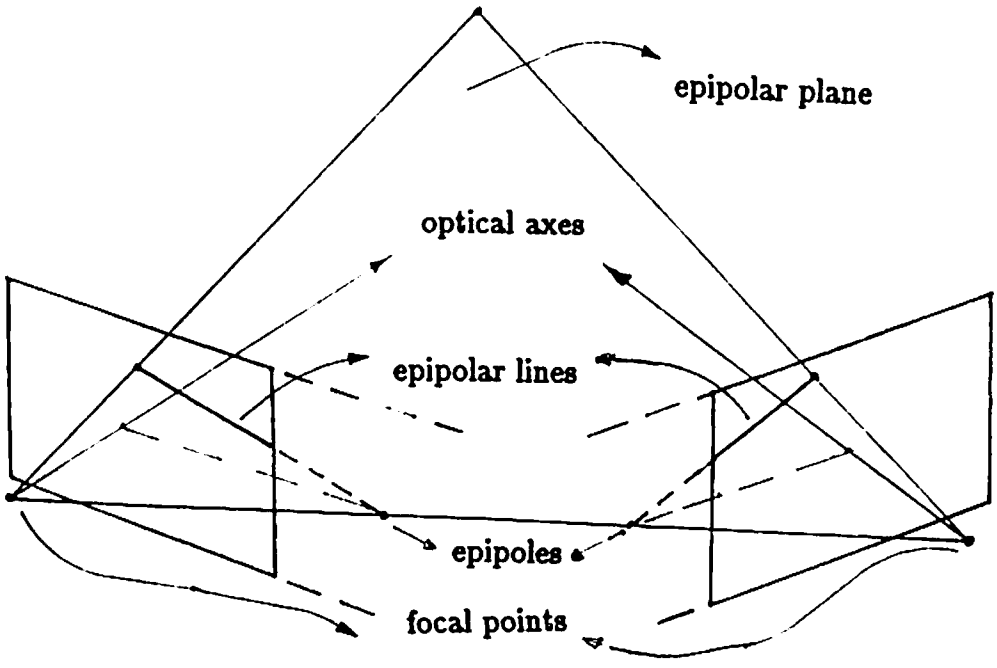

Figure 8: The camera set-up for stereopsis

4.1 The geometry of stereopsis

The geometry of stereopsis is best described in terms of what is known as **epipolar analysis**. This is discussed in detail in [Bake83]. Here, we present some of the major ideas from that paper.

The epipolar analysis considers two cameras placed in arbitrary relative locations and orientations. The camera set-up is shown in figure 8.

The **epipole** is defined to be the point where the line joining the two focal-points intersects an image plane. Therefore there can be at most two epipoles, one for each image. If the line joining the two focal points is parallel to either of the image planes, then there is no epipole for that image.

The two key concepts are the **epipolar plane**, which is the plane containing the two focal points and a point in the environment, and the **epipolar line**, which is the intersection of the epipolar plane with an image plane. It should be obvious that the

image of a point will lie on an epipolar line. Indeed, for each point in an image, the epipolar line can be determined as the line joining that point to the epipole.

Baker, et. al. define a coordinate system for each image based on the location of the epipole. The location of any point on the image-plane can be specified by the epipolar line it lies on, and its distance on that line from the epipole. These are the **epipolar coordinates** of that point.

In the case where the epipole does not exist, it can be shown that there still is an **epipolar direction** and the epipolar lines will all be parallel and oriented along the epipolar direction. Any point on the image-plane can then be arbitrarily chosen as the origin of a cartesian coordinate system whose axes are parallel and perpendicular to the epipolar direction.

If the relative camera locations and their relative orientations are known, then given an epipolar line in one image, the corresponding epipolar line in the other image can be determined. Therefore, for any point in one of the image frames, its corresponding point in the other image must lie along the corresponding epipolar line (called the **conjugate epipolar line**) in the other image. This constraint (called the "epipolar-line constraint") can be used in the matching process.

4.2 Geometric and physical constraints

The geometric constraint often used in stereopsis that the two image-projections of an environmental point lie on conjugate epipolar lines in the two images. This can be used in two ways. One approach involves choosing an event of interest from one image (e.g., an edge, an "interesting point", etc.), determining its epipolar line, and searching for its

match along the conjugate epipolar line in the other image. The other approach involves geometrically transforming the images such that the epipolar lines become horizontal scan-lines. This has some computational advantages, especially when designing an algorithm that operates in parallel at all image-points of interest.

Baker, et al. [Bake83] use the latter approach. Their technique combines an edge-based matching process with an intensity-based correlation algorithm. For the edge-based matching process, they first detect edges from each image and then transform the edge images. For the correlation process, they transform the two intensity images.

In addition to the epipolar-geometry constraints, it is also possible to derive two other physical constraints. In order to describe both constraints, it is convenient to assume that the images have been transformed such that the epipolar lines are scan-lines.

An ordering constraint Consider two points in the environment whose images are on the same scan-line – possibly at different depths. Then, the point which is to the left of the other in the scene will have its images to left of the images of the other in both views. This provides an ordering constraint for the matching process – viz., the left-right relationship of the two points must be preserved across the two views.

A continuity constraint This constraint is similar to the smoothness constraint explained in section 2.3. We reformulate it here in terms of edges in the image.

Edge fragments are usually a part of a contour in the image. Hence, their locations continuously vary across the scan-lines. Therefore, the disparity of the edges that are part of the same contour must vary continuously across scan-lines.

4.2.1 The use of physical constraints

There are a number of algorithms that perform edge-based matching. The ones that are of relevance here are those that use the constraints described above. One such algorithm is provided by Ohta and Kanade [Ohta85].

Their algorithm consists of an intra-scanline search as well as an inter-scanline search. These embody the epipolar line constraint and the continuity constraint. The algorithm also uses the ordering constraint as a part of the intra-scanline search.

In brief, their algorithm uses edge-limited intervals on a scan-line as the events that are matched. An edge-limited interval on a scanline is all the points between successive edges. The algorithm uses a *dynamic programming* approach, which is an optimization technique. This involves minimising a global cost of match, which evaluates simultaneously the matches of all the edge-limited intervals. The match-cost is computed in a systematic manner that enables efficient processing to take place.

Each possible match of edges between a pair conjugate scanlines is given a match-cost determined by the similarity of the intensity values of points in the corresponding edge-limited intervals. The costs from all the scanlines are added together to provide the global measure. The intra-scanline ordering constraint and inter-scanline consistency are used during the computation and combination of these costs. We omit the details of the algorithm here, but simply state that Ohta and Kanade provide a parallel, iterative sceheme for searching for the optimal set of matches.

Baker [Bake82] describes a similar edge-based scanline matching technique using a dynamic programming algorithm. His approach differs from that of Ohta and Kanade in how the measures are added and how the search for the minimum-cost match is performed.

In particular, Baker finds the optimal intra-scanline match independently for each scanline. He then uses a cooperative process that detects and corrects intra-scanline matches that violate the inter-scanline consistency constraint.

4.3 Using domain specific constraints

When the problem is restricted to specific scene domains some additional constraints can be used to partially interpret the static image structures and analyze their transformation across the two views to determine their 3-d structure. We describe here two examples of methods that use such information. Both the examples deal with the stereopsis of urban scenes containing buildings and roads. In such environments there are usually clearly delineable curves that form the boundary between objects and surfaces, and polyhedral vertices which are junctions of physical edges of an object.

4.3.1 The approach used in the 3-d Mosaic system

The 3-d Mosaic system is a vision system that incrementally reconstructs complex 3-d scenes from multiple images. This approach is being developed and studied by Kanade and his colleagues at Carnegie-Mellon University [Herm84].

One aspect of this system is the generation of 3-d information from stereopsis by matching structural features such as junctions of lines. In aerial images, it can be often assumed that L-shaped junctions are the results of surfaces parallel to the ground plane – usually part of building tops, roads, etc. An ARROW or a FORK junction usually arises when three mutually orthogonal lines intersect. These two observations are useful in constraining the search for correspondences.

If the orientation of the image plane is known (or can be estimated) with respect to the ground plane, the location of L-junctions in subsequent frames can be predicted. In the case of the ARROW or FORK, their location as well as their appearance can be predicted (or at least constrained).

In addition, the Mosaic system uses constraints on the relationship between connected junctions. It is assumed that two connected junctions are at the same height from the ground. This assumption is used as a consistency condition for the depth of the two junctions. It is not clear how the system handles violations of this assumption.

4.3.2 The rule based analysis at Stanford

Baker and his colleagues [Bake83] use structural constraints similar to those described above for urban scenes. Such constraints are made part of a rule-based analysis system for the determination of the 3-d structure of objects. Hence, the representation and the manner in which these constraints are used are different from those found in the 3-d Mosaic system.

The analysis is based on structural elements called orthogonal trihedral vertices (OTVs). These are junctions of three mutually orthogonal lines (the same as the ARROW or FORK junctions described above). Based on an analysis of the geometry of OTVs by [Perk68], they provide a method of identifying an OTV in a monocular image and determining its 3-d orientation. The system then matches the OTVs across the strereo pair.

They also derive a set of rules for the 3-d interpretation of image structures (T-junctions, OTVs, and edges) in the monocular image, as well as for the relationship between their views in a stereo-pair. These rules are then used to guide the matching process. These

rules, their representation, and how they are used can be found in [Bake83].

4.4 Stereopsis algorithms based on human vision

Most of the stereopsis algorithms that try to model biological vision are based on the experiments of Julesz that suggest that very simple image-tokens are matched [Jule71]. These experiments are based on *random dot stereograms.* A random dot stereogram is generated from an array of random dots. Typically, a square subsection of an identical copy of the array is displaced a given amount and the resulting gap is filled in with a new random pattern. When the original array and the modified copy are presented to the left and right eyes seperately, one sees a square floating in space This occurs even though there is no monocularly visible square to guide the matching process. Clearly, the matching primitive used in the human visual system is of a very simple nature.

Even though the search for a matching feature in stereopsis is constrained to the epipolar line, the "false target" problem is by no means trivial. Indeed, in the case of a random dot stereogram, any "dot" in the left image could conceivably match any dot along the corresponding raster in the right. Marr and Poggio suggest that the false target problem is solved by exploiting two constraints of the physical world [Marr82]. The "uniqueness" constraint states that each feature in the left image should match at most one feature in the right, since the feature corresponds to a unique point on a physical surface. The "continuity" constraint states that since matter is cohesive and grouped into surfaces, disparity should vary smoothly, except at surface boundaries. (Note that these are examples of the local and global constraints discussed in section 2.3.)

In these approaches, a multiple frequency channel approach – of the kind described

in section 2.4 – is also utilized. Marr and Poggio propose that the features matched in human stereopsis are the zero crossings of the $\nabla^2 G$ convolution with the left and right eye images [Marr79]. This convolution has the effect of applying a band-pass filter to the image. The zero crossings of the filtered image correspond to changes in image intensity at the scale of the associated gaussian. By restricting the search for a matching zero crossing to a sufficiently small interval along the scanline, matches can be determined almost unambiguously. The greater disambiguating power of low spatial frequency tuned channels is in turn exploited by high spatial frequency tuned channels through eye vergence movement.

The biologically oriented stereopsis algorithms also incorporate vergence control as a part of the correspondence process. In an implementation of the Marr and Poggio stereopsis algorithms by Grimson [Grim80], matching is first conducted within the lowest spatial frequency tuned channel. Only zero-crossings are matched, and the only feature associated with them is their sign – i.e., the sign of the intensity variation of the filtered image around the zero crossing when traversed along a particular direction. They are said to have a successful match if a zero-crossing with the same sign is found within a specified disparity range. Matching is then attempted at the next higher frequency. Any local area of the higher frequency channel with less than 70% successful matches is declared "out of range". Regions that are out of range require eye vergence movement, a relative adjustment of the position between corresponding local areas of the left and right eye image. In order to determine in which direction the local area should be shifted, the "majority disparity" [4]

[4]The process of determining the "majority disparity" involves histogramming, where each point in the area votes for its disparity. The disparity value with the maximum number of votes is selected as the "majority disparity"

within the corresponding area of the low frequency tuned channel is computed. Depending on whether the majority of the disparities lie to the left or to the right, the local area is shifted in the appropriate direction and matching is repeated. The process is repeated until all local areas within all spatial frequency tuned channels have been processed. Grimson's program has been demonstrated on several random dot stereograms and natural images with good results.

There are several problems with the Grimson control strategy. The most significant is its failure on images with periodic features [Grim80]. Since a local area of a high spatial frequency tuned channel only requests vergence movement when it is locally dissatisfied, and any one of a number of possible alignments will satisfy it in an image with periodic properties, different initial vergence positions produce different results. No attempt is made to reconcile the local definition of disparity with conflicting estimates provided by more global sources (i.e., those provided by neighboring areas or other channels). Additionally, since vergence is realized physically by eye movement, and since different local areas can simultaneously make conflicting requests of the eye movement resource, questions have been raised about its adequacy as a human model [Will85].

Mayhew and Frisby propose that computation of correspondence in stereopsis is closely linked to the construction of a symbolic description of image intensity changes occuring at different spatial scales, called the "raw primal sketch" [Mayh81]. Central to Mayhew and Frisby's theory is the notion of "spectral continuity" proposed by Marr and Hildreth [Marr80]. This is similar to the ideas described in 2.4. The spectral continuity constraint states that disparity of the primal sketch token should remain relatively constant over a range of spatial frequencies. Mayhew and Frisby suggest that matches that preserve

spectral continuity and "figural" continuity (unbroken zero crossing contours) should be selected over matches that do not.

The notion of spectral continuity has recently been incorporated in a vergence strategy that attempts to address some of the problems inherent in Grimson's control strategy [Will85]. In Williams' implementation, the left and right images are moved relative to each other in a single uniform movement. Matching is repeated at periodic intervals; matches within the higher frequency channel are accepted only if they agree with the "majority disparity" within the corresponding area of the next lower frequency. Matching within a local area of a spatial frequency tuned channel is viewed as taking opportunistic advantage of a vergence movement that is controlled at a global level.

4.5 Obtaining 3-d descriptions

Most of the correspondence techniques provide only a depth map, i.e., a specification of the depth of each point in the image. In a practical system, it may be more useful to obtain information about 3-d surfaces – their shape, extent, location and orientation in space, or a volumetric description of the objects in the environment.

The following review of 3-d representations and how they relate to stereopsis is based on an excellent overview of 3-d object recognition provided by Paul Besl and Ramesh Jain [Besl85].

The major categories of object representations are the *wire-frame representation*, which consists of vertices and connecting edges, the *constructive solid geometry description*, which specifies objects as a combination of volumetric primitives such as cones and cylinders, the *spatial occupancy representation*, which specifies the space occupied by a particular object,

and the *surface boundary representation*, which mathematically specifies the shape of the surfaces. In some cases, it is convenient to represent the object in terms of its appearance from a set of views.

The issue of how to obtain these models from the depth information is an open problem. Besl and Jain survey a number of efforts to convert depth map information to one or another of these models. There have been attempts to segment depth-maps, locate discontinuities in the depth map, locate 3-d edges and junctions from depth maps, and describe depth-map segments as planar or other polynomial surfaces. A complete survey and bibliography can be found in [Besl85].

4.6 A few remarks

In certain ways stereopsis is a simpler problem than motion, since the relative location and the orientation of the cameras is under the control of the user, or can be predetermined. This knowledge provides additional constraints that simplify the search for the correspondence of image events. However, even in this simpler case, there are few attempts to extract 3-d information from the depth maps and to integrate depth information in practical real-time systems.

5 CONCLUSION

This chapter has provided an overview of some of the issues involved in stereopsis and motion analysis in computer vision. In this section, we summarize the major aspects of contemporary approaches, and discuss possible next steps in the development of these areas of research.

The major aspects of contemporary approaches are:

1. The important results in both motion and stereopsis have contributed towards a better understanding of the geometry of disparity and optic flow fields.

2. The major problem in stereopsis is the correspondence problem, since the knowledge of the relative camera geometries simplifies the interpretation of disparity data.

3. In motion, interpreting the low-level correspondence data is an equally complex problem. The emphasis has been largely on obtaining accurate quantitative results regarding 3-d structure and motion. However, most of the techniques appear to be unstable and incapable of handling a wide variety of imaging situations and different types of motion.

4. In both stereopsis and motion the correspondence problem has been largely addressed in terms of low-level image data such as intensity variation, points with stable image properties, edges, and lines. Although in motion analysis some attempts have been made, to use larger and more complex image structures, the dependence of such methods on good static processes has handicapped their usage.

5. Most of the motion-correspondence algorithms simply find the displacement or 2-d velocity of a point, or its average over an area. Not much has been done to use the change in shape of curves and regions in the image.

6. Both motion and stereopsis work has concentrated on describing the 3-d structure of the scene by specifying the depth of the image points. Very little work has been done to extract 3-d surface or volumetric decriptions.

7. There is virtually no work that integrates motion and stereopsis. This seems peculiar, since similar techniques are used in both problems.

8. Almost all of the motion work so far has concentrated on analyzing the information from a pair of frames. Expanding this work to process a sequence of more than two images is not a simple task.

9. Almost no attempt has been made to use higher-level control strategies to focus the attention of the camera on locations of interest. Most of the processing proceeds in a bottom-up fashion, from image, to motion-data, to its interpretation.

These comments focus mainly on the shortcomings of current approaches. This was done in order to indicate the magnitude of the task ahead. In what follows, we briefly discuss two issues – integrating motion and stereopsis, and processing a longer sequence of images.

5.1 Integration of motion and stereopsis

Integration of motion and stereopsis arises when two cameras in a stereo configuration move together. In his recent work, Jenkin [Jenk84] describes some possible approaches to this problem.

At the level of the correspondence problem, the integration process can proceed in one of three possible ways. Stereopsis can be performed before temporal matching, motion analysis can precede stereopsis, or both can be done simultaneously. Jenkin chooses the third approach in order to let the two processes aid each other.

His approach consists of the cyclical operation of four successive modules called static

analysis, prediction, testing and decision. Static analysis extracts monocular image features (usually feature points of the type discussed in this chapter) and lists all potential binocular matches of these features. The prediction module uses the static analysis as well as the motion information from the previous frame to predict and constrain the motion of the features. A set of hypotheses concerning the motion of the features is derived. The testing module uses the information from the next frame to identify invalid hypotheses. The remaining ones are used by the decision module to update the scene model and the motion information. Details can be found in [Jenk84].

5.2 Processing longer image sequences

There are not many examples of systems that involve processing more than two frames at a time. Some of the correspondence algorithms can be extended to use more than two frames in some simple way. For example, the gradient-based algorithms require the temporal derivative of the image intensity variations at a point. This can be obtained in some manner from a sequence of frames. Fleet [Flee84] proposes a model for a velocity detection mechanism based on neuro-physiological data concerning cortical cells.

The matching algorithms can be extended by using a temporal form of the smoothness constraint on the optic flow. Some discussion of this can be found in [Horn80] and [Prag83]. Although these ideas are somewhat old, it is interesting to note that there is no system that actually incorporates any of them.

Jenkin's work described above [Jenk84] also includes processing a sequence of images using a prediction and verification mechanism. Another prediction-based algorithm for processing images generated by the translation of a camera is also proposed in [Bhar85].

These ideas are as yet preliminary. This area of research is likely to be the next important development in motion analysis.

References

[Adel83] Adelson E. H. and Movshon J. A. The Perception of Coherent Motion in Two-Dimensional Patterns *ACM Workshop on Motion*, Toronto, Canada, pp. 11-16, 1983.

[Adiv85] Adiv G., Interpreting Optical Flow, *Ph. D. dissertation*, COINS Department, University of Massachusetts, Amherst, September 1985.

[Anan84] Anandan P., Computing Dense Displacement Fields with Confidence Measures in Scenes Containing Occlusion, *SPIE Intelligent Robots and Computer Vision Conference*, Vol. 521, pp. 184-194, 1984, also *COINS Technical Report 84-32*, University of Massachusetts, December 1984.

[Anan85] Anandan P. and Weiss R., Introducing a Smoothness Constraint in a Matching Approach for the Computation of Optical Flow Fields, *Proceedings of the Third Workshop on Computer Vision*, Michigan, October 1985, pp. 186-194.

[Bake82] Baker H. H., Depth from Edge and Intensity Based Stereo, *Report No. STAN-CS-82-930*, Department of Computer Science, Stanford University, California, September 1982.

[Bake83] Baker H. H., Binford T. J., Malik J., and Meller J., Progress in Stereo Matching, *Proceedings of DARPA IU Workshop*, Virginia, June 1983, pp. 327-335.

[Ball82] Ballard D. H. and Brown C. M., *Computer Vision*, Prentice-Hall Inc., New Jersey, 1982.

[Barn80] Barnard, S. T. and Thompson, W. B., Disparity Analysis of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-2, Number 4, July 1980, pp. 333-340.

[Besl85] Besl P. and Jain R., Range Image Understanding, IEEE CVPR Conference Proceedings, June 1985, pp. 430-449.

[Bhar85] Bharwani S., Riseman E. M., and Hanson A., Refinement of Environmental Depth Maps Over Multiple Frames, *Proc. DARPA IU Workshop*, 1985, in press.

[Burt82] Burt, P. J., Yen, C. and Xu, X., Local Correlation Measures for Motion Analysis: A Comparative Study, *IEEE Proceedings of PRIP*, 1982, pp. 269-274. Also *IPL-TR-024*, ECSE Dept., RPI, 1982.

[Burt83] Burt, P. J., Yen C. and Xu X., Multi-Resolution Flow-Through Motion Analysis, *IEEE CVPR Conference Proceedings*, June 1983, pp. 246-252.

[Corn83] Cornelius N. and Kanade T., Analysing Optical-Flow to Measure Object Motion in Reflectance and X-ray Image Sequences *ACM Workshop on Motion*, Toronto, Canada, April 1983, pp. 50-58.

[Duda73] Duda R. O. and Hart P. E., Pattern Recognition and Scene Analysis *Wiley*, New York, 1973.

[Dres81] Dreschler L. and Nagel H. H., Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene, *Computer Graphics and Image Processing*, 20 (3), pp 199-228, 1982.

[Fenn79] Fennema C. L. and Thompson W. B. Velocity Determination in Scenes Containing Several Moving Objects, *Computer Graphics and Image Processing*, 9, pp 301-315, 1979.

[Flee84] Fleet D. J., The Early Processing of Spatio-Temporal Visual Information, Tech. Report No. RBCV-TR-84-7, University of Toronto, September 1984.

[Genn80] Gennery D. B., Modelling the Environment of an Exploring Vehicle by Stereo Vision, *Ph. D. thesis*, Stanford AI Laboratory, June 1980.

[Glaz81] Glazer F., Computing Optic Flow, *IJCAI-7*, Vancouver B. C., Canada, Aug. 1981, pp. 644-647.

[Glaz83] Glazer, F., Reynolds, G. and Anandan, P., Scene Matching by Hierarchical Correlation, *IEEE CVPR conference*, June 1983, pp. 432-441.

[Grim80] Grimson W. E. L. A Computer Implementation of a Theory of Human Stereo Vision, *Philosophical Tranactions of the Royal Society of London*, vol. B292, pp. 217-253.

[Hann74] Hannah, M. J., Computer Matching of Areas in Stereo Images, *Stanford A.I. Memo 239*, July 1974.

[Herm84] Herman M. and Kanade T., The 3D MOSAIC Scene Understanding System:

Incremental Reconstruction of 3D Scenes from Complex Images, *Proceedings of the DARPA IU Workshop*, October 1984, pp. 137-148.

[Hild83] Hildreth, E. C., The Measurement of Visual Motion, *PhD dissertation*, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, Ma., 1983.

[Horn80] Horn, B. K. P. and Schunck, B. G., Determining Optical Flow, MIT A.I. Memo Number 572, April 1980.

[Humm83] Hummel R. A. and Zucker S.W., On the Foundations of Relaxation Labeling Processes *IEEE PAMI*, vol 5, no. 3, 1983, pp. 267-286.

[Jenk84] Jenkin M., The Stereopsis of Time-Varying Images, *RBVC Technical Report, RBVC-TR-84-3*, Dept. of Computer Science, Univ. of Toronto, Ontario, Canada, 1984.

[Jule71] Julesz B., Foundations of Cyclopean Perception, *University of Chicago Press*, Chicago, 1971.

[Kitc80] Kitchen L. and Rosenfeld A., Grey Level Corner Detection *Tech. Rep. No. 887* Computer Science Center, Univ. of Maryland, College Park, 1980.

[Lawt84] Lawton D. T., Processing Dynamic Image Sequences from A Moving Sensor, *PhD dissertation*, COINS Dept., Univ. of Massachusetts, TR 84-05, 1984.

[Limb75] Limb J.O. and Murphy J. A., Estimating the Velocity of Moving Images in Television Signals, *Computer Graphics and Image Processing*, vol. 4, 1975, pp. 311-327.

[Long80] Longuet-Higgins H. C. and Prazdny K., The Interpretation of a Moving Retinal Image, *Proceedings of the Royal Soc. London*, B208, 1980, pp. 385-397.

[Marr79] Marr D. and Poggio T., A Computational Theory of Human Stereo Vision, *Proc. Roy. Soc. London*, B204, pp 301-308, 1979.

[Marr80] Marr D. and Hildreth E., Theory of Edge Detection *Proc. Roy. Soc. London*, B207, pp. 187-217.

[Marr81] Marr D. and Ullman S., Directional Selectivity and its Use in Early Visual Processing, *Proc. Roy. Soc. London*, B211, 1981, pp. 151-180.

[Marr82] Marr D. *Vision*, W. H. Freeman and Co., San Francisco, 1982.

[Mayh81] Mayhew J. E. W. and Frisby J. P. Psychophysical and Computational Studies Towards a Theory of Human Stereopsis, *Artificial Intelligence*, 17, 1981, pp. 349-385.

[Mora80] Moravec H. P., Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover, *Ph. D. thesis*, Stanford University AI Laboratory, California, September 1980.

[Nage83a] Nagel H. H., Displacement Vectors Derived from Second-Order Intensity Variations in Image Sequences, *Computer Vision, Graphics, and Image Processing*, 21, pp 85-117, 1983.

[Nage83b] Nagel H. H., Constraints for the Estimation of Displacement Vector Fields from Image Sequences, *IJCAI-83*, Karlsruhe, W. Germany, pp 945-951, 1983.

[Nish84] Nishihara K. Practical Real-Time Imaging Stereo Matcher, *Optical Engineering*, 23 (5), pp 536-545, 1984.

[Ohta85] Ohta Y. and Kanade T., Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming, *IEEE PAMI*, vol. 7, no. 2, pp. 139-154.

[Perk68] Perkins, Cubic Corners,*Quarterly Progress Report*, MIT Electronics Lab., 1968.

[Prag83] Prager J. M. and Arbib M. A. Computing the Optic Flow: The MATCH Algorithm and Prediction, *Computer Vision, Graphics, and Image Processing*, 24, pp 271-304, 1983.

[Quam84] Quam L. H., Hierarchical Warp Stereo, *Proceedings of DARPA IU Workshop*, Louisiana, October 1984, pp. 149-156.

[Rieg83] Rieger J. H. and Lawton D. T. Determining the Instantaneous Axis of Translation from Optic Flow Generated by Arbitrary Sensor Motion, *Proceedings of the ACM Workshop on Motion*, Toronto, Canada, 1983, pp. 33-41.

[Roac80] Roach J. W. and Aggarwal J. K., Determining the Movement of Objects from a Sequence of Images, *IEEE PAMI*, vol. 2, no. 6, 190, pp. 554-562.

[Thom81] Thompson, W. B. and Barnard, S. T., Lower-Level Estimation and Interpretation of Visual Motion, *Computer*, Aug. 1981.

[Tsai84] Tsai R. Y. and Huang T. S., Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces, *IEEE PAMI*, 6, 1984.

[Waxm83] Waxman A. M. and Ullman S., Surface Structure and 3-d Motion from Image Flow: A Kinematic Analysis, *Research Tech. Rep. No. 24*, Center for Automation, Univ. of Maryland, October 1983.

[Waxm84a] Waxman A. M., An Image Flow Paradigm, *Proceedings of the Workshop on Computer Vision*, 1984, pp. 49-57.

[Waxm84b] Waxman A. M. and Wohn K., Contour Evaluation, Neighbourhood Deformation and Global Image Flow: Planar Surfaces in Motion, *CS-TR-1394* University of Maryland, April 1984.

[Will85] Williams L. R., Spectral Continuity and Eye Vergence Movement, *Proceedings of the ninth IJCAI*, California, 1985, pp. 985-987.