A Computational Study of Rigid Motion Perception

Amit Bandopadhay
Department of Computer Science
The University of Rochester
Rochester, NY 14627

TR221
December 1986

This report reproduces a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The work was supervised by Dr. Dana H. Ballard.

This work was partially supported by the National Science Foundation under grant DCR-8405720 and the Office of Naval Research under contracts N00014-80-C-0197 and N00014-82-K-0193.

We thank the Xerox Corporation University Grants Program for providing the equipment used in the preparation of this paper.

Curriculum Vitae

Amit Bandopadhay was born in Calcutta, India in December 1954. He had his schooling at various suburban towns in the State of West Bengal and at Calcutta. In 1972 he graduated high school and was the recipient of the Indian Government's National Merit Scholership. In the same year he joined the Indian Institute of Technology (IIT) at Kharagpur, to study electronics and electrical engineering, obtaining the B.Tech. degree with honours in 1977. At Kharagpur he was involved in a wide variety of activities such as dramatics, debating, music for which he received many awards. He served as the Governor of the Technology Dramatic Society for a year. He is also an avid hiker and enthusiastic sports lover. In 1977 he joined the Indian Institute of Technology at Kanpur from where he graduated in 1979 with the M.Tech. degree in computer science. While at Kanpur he was elected as the president of the Association for Computing Activities.

The next phase of his life was spent in Calcutta, from 1979 to 1981, working as a systems analyst for a multinational corporation.

He joined the Department of Computer Science at the University of Rochester in the fall of 1981. The field of artificial intelligence had always fascinated him so he decided to work in computer vision. He was fortunate to have found a stimulating atmosphere in the Department of Computer Science, and was enriched by the intellectual atmosphere there. His doctoral thesis was supervised by Professor Dana Ballard. While at the University of Rochester he served as teaching assistant and research assistant from the fall of 1981 to the summer of 1986.

He has published about a dozen papers and technical reports, in robot motion perception. His other interests are in massively parallel computations, knowledge representation and retrieval, expert systems, automated learning, very large databases, evidential reasoning and computer aided instruction. He is a member of the Association of Computing Machinery (ACM) and the Society of Photo-Optical Instrumentation Engineers (SPIE).

Acknowledgements

It is hard to even begin to thank adequately, the many people who have influenced my thinking and stimulated my intellectual growth.

Dana Ballard, my thesis supervisor, has provided invaluable help at all stages of this work. He showed me how to separate the wheat from the chaff. When there were dry spells, and ideas were not forthcoming, his indefatigable enthusiasm kept me going. You have my thanks and admiration for your judicious support, patience and for your help in all the tangible and intangible ways that are difficult to enumerate.

I consider myself fortunate to have had the chance to interact with Jerry Feldman. His clear vision has helped me to put things in perspective and, I hope, will give me courage to break out of established dogma when necessary. My discussions with him have invariably proved to be valuable in expanding my comprehension of my chosen field.

Chris Brown is a guiding spirit of the Rochester Vision Group. His energy and leadership invigorates the vision research done here. In spite of his numerous duties as the chairman of the Computer Science Department, he has almost never failed to find time for discussions or advice, whenever I

or any other student needed it.

I thank Peter Lennie for agreeing to sit in my thesis committee, and for tolerating proposals for solving vision tasks that surely seemed, at times, far removed from the scheme of Nature.

My stay at Rochester has been enriched by the friends I have had the fortune of having here. Yannis Aloimonos and Vally Koubi were always there for support and help. Yannis has always been an abundant source for technical inspiration. The vision group deserves thanks for making this place a rewarding place to work in. I also wish to thank Paul Chou, Joel Krenis, Takahide Ohkami, Doug Ierardi for their friendship. Barun Chandra and Rabi Dutta, were great people to work with. Thanks to Gary Cottrell, Hari Narayanan, Cesar Quiroz, Ken Yap, Isidore Rigoutsos, Rich Pelavin, Richard Newman-Wolfe and Josh Tenenburg for companionship.

I would like to remember Lydia Hrechanyk, who had a rare combination of intelligence, charm and patience. She affected us, felow graduate students and office mates, in more ways than can be recounted.

I am happy to also thank Mr. and Mrs. Ronald Furman, for friendship and caring, especially during the first few days of my stay here. Thanks also to Tamisra Sanyal, Atul Kacker and Amitabha Mukherji, long suffering apartment mates past and present.

Finally I thank my parents for instilling in me the desire to know, and my siblings for loving without expecting any return.

Abstract

The interpretation of visual motion is investigated. The task of motion perception is divided into two major subtasks: (i) estimation of two dimensional retinal motion, and (ii) computation of parameters of rigid motion from retinal motion. Retinal motion estimation is performed using a point matching algorithm based on local similarity of matches and a global clustering strategy. The clustering technique unifies the notion of matching and motion segmentation and provides an insight into the complexity of the matching and segmentation process. The constraints governing the computation of the rigid motion parameters from retinal motion are investigated. The emphasis is on determining the possible ambiguity of interpretation and how to remove them. This theoretical analysis forms the basis of a set of algorithms for computing structure and three dimensional motion parameters from retinal displacements. The algorithms are experimentally evaluated. The main difficulties facing the computation are seen to be nonlinearity and a high dimensional search space of solutions. To alleviate these difficulties an active tracking method is proposed. This is a closed loop system for evaluating the motion parameters. It is shown that

under such a regime it is possible to obtain closed form solutions for the motion parameters. This leads to a robust cooperative algorithm for motion perception requiring minimal amount of retinal motion matching. The central theme for this research has been the evaluation of a hierarchical model for visual motion perception. To this end, the investigations revolved around three primary issues: (a) retinal motion computation from intensity images; (b) the conditions under which three dimensional motion may be computed from retinal motion, and the efficacy of algorithms that perform such computation; (c) the active vision or closed loop approach to visual motion interpretation and what it buys us. This thesis records fundamental contributions pertaining to the above questions.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

1.1. The Motion Perception Problem

Our visual perception creates awareness of the world around us, that consists of a rich variety of objects. These objects are characterized by different shapes, colors and motion patterns. The visual data (or stimulus) that is captured by the eyes is in essence two dimensional patterns of light reflected from the surfaces, normally of solid rigid objects, that exist in our environment. When we ponder the complexity of the three dimensional scene surrounding us, it is apparent that despite the unconscious ease with which our brain interprets the visual data available to it, Visual Perception is a complicated task. Two of the problems associated with actually "seeing" in three dimensions are immediately apparent. First, the images formed on the retina of the eyes are two dimensional, thus three dimensional information is only implicit. Second, the retinal images are continually changing, due to the movement of objects we are seeing, or due to our own movement.

To relate retinal images to object models natural constraints in the world must be used. The retinal stimulus contains implicit information that can be used to recover aspects of the three dimensional world being viewed. There are two images formed on the retinas of the two eyes, which are spatially displaced from each other. The principle of stereoscopic fusion (or triangulation) of the images of a point in space to compute its depth has been recognized for a long time. There is also the intimate relationship between the local surface shape (e.g. slant and tilt) and motion induced change in the retinal intensity pattern.

This thesis is concerned with one particular set of visual constraints, namely those having to do with motion. The problem under study concerns the task of computation of the three dimensional motion between an observer and rigid objects.

It is now widely accepted that motion is a fundamental sense or modality, that is extracted from the visual stimulus array (see [63]). This computational study of motion will be restricted to stimuli that contain information primarily about spatio-temporal variations in the image intensity distribution. However, it is useful to bear in mind that submodalities like depth and surface orientation can prove helpful in analyzing the motion understanding process. On the other hand color information is assumed to play a minimal role in the perception of motion. Therefore, the visual input that is considered useful is monochromatic images from either one or both

eyes.

The computational goal of the motion perception process is to obtain estimates for the three dimensional velocities of the objects being observed. The latter quantities are also termed relative *motion parameters* and are global attributes of the moving bodies. In a theoretical sense there is not much difference in analyzing a scene containing a single moving object and another containing multiple objects in motion. In the latter case, the motion analysis must first perform segmentation or break up the two dimensional image into the various regions corresponding to the different object surfaces. Subsequent to this, the individual segments can be treated separately as image fragments dealing with single body motion.

In subsequent portions of this document, unless otherwise specified, the treatment of three dimensional motion interpretation deals with *egomotiotL* This is the situation where the motion stimuli are generated due to the movement of the observing system in a static visual environment. The reasons for this simplification are

(i) The two dimensional motion estimation algorithm that is proposed in chapter two can handle motion segmentation and hence subsequent analysis need not deal with more than one moving surface.

(ii) Mathematically, there is no difference between the motion stimuli due to a static observer, whose entire visual field registers motion due to one moving object, and that for a moving observer registering the

relative motion of the static surround.

The problem addressed in this thesis can thus be stated as:

> *Problem Definition: Given* monocular or binocular spatio-temporally varying images and known viewing parameters, *to compute* egomotion parameters and structure of the imaged scene.

The viewing parameters referred to in the above definition are the focal length, image scaling factors and the relative locations and orientations of the two cameras (in case of binocular imagery).

Traditional approaches to this problem have made two different kinds of assumptions when compared to the methodology advocated here. The first of these is that the monocular stimulus should be enough to compute the motion parameters. The theoretical basis of this belief will be explored to evaluate how well monocular data can be used to aid the perception process. It will be seen that the problem is beset with two principal difficulties, namely: (i) *nonlinearity* and (ii) *high parameter space dimension*

The above difficulties make computer algorithms for motion computation complex and sensitive to errors in two dimensional retinal motion measurement [84].

Almost all previous work is based on the assumption that the motion problem can be solved with passive observation. In passive observation, the sensors (cameras) are rigidly attached to the body in motion. Since motion

is relative, one can always assume that the sensor is fixed and the environment moves. Thus *passive navigation* deals with the measurement of motion with respect to a static sensing system. This problem assumes that the alignment of the camera axis with respect to the three dimensional velocity directions is arbitrary and fixed. In general the solution can be shown to be dependent upon nonlinear equations of large dimensions [19, 64]. There is no reason to believe that passive navigation can lead to efficient and robust solutions to the problem of motion perception. In fact it will be shown that such methods have inherent ambiguities in so far as motion interpretation from two dimensional retinal cues is concerned.

An alternative approach to the problem, can be based upon the assumption that the alignment of the camera axes are controllable by the observer. In this case, as the observer continues to move in the world, the orientations of the eyes (cameras) are continually adjusted. This adjustment is dependent upon the two dimensional motion perceived on the retina, and serves - among other things which will be explained later - to simplify the constraints governing the perception of the motion. This is the mechanism of *active navigation* that will be explored subsequently.

The goal of this dissertation is to explore computational solutions to the problem of rigid motion perception. A key orientation of this research has been to derive inspiration for the structure of the computer model from relevant known attributes of the primate visual system. This knowledge,

together with some of the emerging concepts in computer and cognitive science regarding highly parallel computational models [31, 32] and parameter estimation and transformation [7, 9, 17, 18] motivated the proposed computational scheme. Before elaborating on this model some aspects of the biological vision mechanisms are examined*

1.2. The Primate Visual System.

This section examines some interesting attributes of biological vision. The account is not meant to present a comprehensive picture of neural visual processing. Rather, the aim is to highlight important neurobiological features that have strong computational advantages, and form an important motivation for the motion model proposed in this thesis.

1.2.1. Abstraction Hierarchies

The cortex, which is the outermost portion of the brain, can be roughly regarded as a two dimensional sheet, a few millimeters in thickness. This sheet consists of gray matter, which are the neuronal computing units and white matter, which constitute the mass of fibers that the neuronal units use to communicate with each other.

Neuroscientists have been able to partition the cortex into a number of distinct areas. The notable property that emerges is that of uniformity of the processing architecture, coupled with the functional diversity of the different areas [54]. The primary visual areas in the *striate* cortex are

retinotopic. This means that they encode information in the visual field indexed by two dimensional retinal coordinates. Thus for example a bright spot of light shone at a particular angular position in the visual field will affect only those units that are responsible for the given retinal position.

There is good evidence to suggest that different cortical areas compute and represent information *at different levels of abstraction* [26]. An indication of this is provided by an experiment by Movshon [59], which compared the responses of neurons in areas V1 and MT in the macaque monkey. Given a checkerboard stimulus, neurons in V1 responded optimally when the motion was perpendicular to the intensity gradients of the checkers. This behavior is isotropic with respect to the orientation of the intensity gradient and only depends upon the magnitude of the temporal intensity change, which is maximum when the motion is perpendicular to the intensity gradient of the checkers. On the other hand when some of the neurons in the MT were probed, the responses indicated that each had its own preferred direction of motion.

An interpretation of the above could be that the V1 neurons are involved in the computation of temporal change in image intensity, while the MT units code *optical flow,* which is the retinotopic projection of the three dimensional velocity field. Such indication of different abstraction levels were first observed by Hubel and Wiesel [46, 47], who postulated a hierarchical functional architecture for visual processing with successively

more and more abstract neuronal units which they called *simple, complex* and *hypercomplex* cells respectively.

In the context of motion perception, one can think of parameters that are at higher levels of abstraction. For instance, in the case of rigid body motion, an economical representation is provided by global (i.e. non retinotopic), parameters such as translation and rotation. In fact, Sakata [74] has identified neurons that respond to full-field rotations, in the *parietal cortex.*

It seems that there exists a motion processing hierarchy in the primate brain. This information processing pathway includes the primary visual cortex (area VI), the middle temporal visual area (MT), the medial superior temporal visual area (MST), and the parietal cortex (area &a) [26]. The parietal cortex and area MST are layers in the motion hierarchy that appear to compute high level motion features. While, the area MT seems to compute lower level retinotopic (i.e. two dimensional) motion representations.

The foregoing discussion highlights some important *design methodologies* in the biological hardware. These have **to do** with *massive* ***parallelism, computation in hierarchies*** **and successive *invariance levels* characterized by** their own parameter sets [10]. The lessons drawn from these attributes will be elaborated subsequently. However, before doing that we will take a look at an important control principle in the motion processing system.

1.2.2. Smooth Pursuit Eye Movements

While it is true that not all creatures with eyes are able to move them, those that do, do so in order to see better. One of the problems with visual perception by a moving observer is that the motion induces blurring. As a quantitative estimate one may calculate that a target movement as slow as 1^0/s, when any point on the target takes about three minutes to cross the visual field, has roughly the same effect on resolution as three diopters of myopia [93]. Thus it is readily seen that one of the reasons for the ability of the eyes to rotate with respect to the head is to stabilize the moving retinal image.

This method of compensation has its limitations, however, since the eyes cannot displace themselves with respect to the head up to any significant degree. Therefore, since the rotational movement has its limits, there are two types of eye movements, both rotational. The first is called *optokinesis* or *smooth pursuit*. This is a relatively slow and continuous movement, whereby the image of a small target can be held steady on the central part of the retina. This is the tracking movement we are primarily interested in. The second type of movement is called a *saccade*, whereby the eyes execute a 'catching' movement to position the image of the object on the central part of the retina. The velocity with which this movement is executed is quite large, being of the order of 1000^0/s [72].

For a human, the pattern of eye movements is likely to be a sequence of saccades with smooth pursuit movement in between. This pattern is called *optokinetic nystagmus.* Consider for instance, a jogger running at a relatively steady pace. His eyes are continually moving according to the following steady pattern ([22]):

(1) A target feature is selected in the environment.

(2) A saccade is made to catch the target and align its image with the optical axis.

(3) A smooth pursuit movement of the eyes takes place, where the target is tracked and held steady.(i.e. the retinal slip is kept as near to null as possible)

(4) When the rotational displacement of the eyes, reaches some limit another suitable target is selected at the periphery of the visual field and steps (2) to (4) are repeated.

The above behavior pattern, seems to support the claim, put forth by Cutting, that the pursuit system plays a cardinal role in our ability to navigate in a cluttered environment.

The most interesting aspect of the smooth pursuit or tracking system is that it illustrates an *active* principle in human visual processing. In other words, the system is closed loop. This point is worthy of reiteration, since it is eminently sensible, even from a system theoretic point of view to to

design measurement mechanisms that adapt to the changes in stimulus. It will be shown subsequently that there are good reasons for designing computer models for motion perception in a similar manner.

Experimental evidence indicates that the primate pursuit mechanism works best when the retinal target velocity is not more than 30°/s [73]. Two other quantitative performance parameters of this mechanism that are relevant are the information processing latency within the control loop, which is around 100 ms. and the tracking error which is found to be well within 10 percent of the target velocity.

In summary, it should be mentioned that an active method forms a dominant principle in the motion perception scheme in primates, and furthermore:

(i) The pursuit hardware is an integral part of the motion processing pathways. (Recall Cutting's observations on how tracking facilitates navigation).

(ii) The selection of the target to be tracked depends on image features such as luminance, size of target (smaller the better), position in the visual field (smaller eccentricities preferred) and velocity. Although small punctate targets are preferred, humans can take advantage of aggregate motion to pursue targets that are perceivable but not visible. An example is the ability to track the center of a rolling wheel that is marked only by several, small lamps attached to its rim [77].

(iii) The 100 ms loop latency seems to be divided into two time steps. The first step, of about 40 ms duration, is distinguished by the fact that the system seems to consider only the direction of target motion for its computations but not the speed.

1.3. Computer Vision, Connectionism, Biology and Computational Structure

The study of machine vision systems cannot ignore the fact that most of the tasks that one sets out to solve are modeled after biological vision. Studies of the human information processing system are affecting the design of machine models for similar tasks in most radical ways, as computer and cognitive scientists are increasingly becoming aware of the fact that conventional stored program concepts inhibit the formulation of cognitive tasks in a fast, robust, adaptive, fault tolerant manner [31].

The foregoing sketch of some aspects of the primate visual system has served to provide a rationale for us to inquire whether it is a good idea to incorporate concepts from Nature into computer models. The specific task at hand is motion interpretation. This section will discuss the design decisions that were made regarding the structure of the proposed computer model for motion perception.

The study of computer vision is conventionally divided into three levels:

(i) *Low Level Vision:* This stage is concerned with early processing of visual information. This level is characterized by the local nature of the computations performed. In the vision mechanism of primates, this stage encompasses the visual processing at the retina and continued till the primary visual cortex* In computer vision, examples are provided by computations connected with the formation of the primal sketch representation of Marr [57] or Feldman's retinotopic frame [32]. Operations at this level are exemplified by filtering, convolution and relaxation based on local constraints.

(ii) *Intermediate Level Vision:* This level of processing is characterized by two major endeavors, namely segmentation and the computations of parameters that signal regional invariance characteristics. Here the main task is to compute stimulus representations that will be used in the next level of visual tasks. One characteristic of encodings at this level is that they are *intrinsic* properties of the viewed objects, and are independent of particular viewing conditions. Examples of representations at this level are, optical flow, field of surface normals corresponding to visible surfaces [15]. This level is also exemplified by Marr's $2\frac{1}{2}D$ sketch and Feldman's stable feature frame. Segmentation is an operation that is quite crucial at this level, because of the need to separate out image regions corresponding to different objects or moving surfaces. This separation is, invariably, a difficult task but serves to simplify higher

levels of computation by preventing effects of independent phenomena from interfering with each other in the analysis (this latter problem is sometimes referred to as *crosstalk* in connectionist literature). Segmentation and parameter estimation at this stage often involve interacting goals* It is likely that cooperative computational algorithms and constraints derived from higher level computational layers are likely to facilitate the processing at this stage.

(iii) *High Level Vision:* This is the level of symbolic information processing. The central task at this level is concerned with, what has been called the *indexing problem* [30]. This is the problem of deriving the description of a situation from a set of visual features computed at the lower levels. At this level methods for knowledge representation, storage and retrieval are of crucial importance. An example of a model of computations can be found in [32], where a dynamically modified store of objects and relations in the observer's extrapersonal space called an *environmental frame,* interacts with a more permanent repository of world knowledge called the *world knowledge formulary.*

The task addressed by this thesis spans the first two levels of the above hierarchy. The computation of motion parameters begins from a time varying sequence of images. One can imagine this input to be akin to a number of consecutive frames of a movie or video sequence. Clearly, there is a difference between the natural visual input to our eyes and this spatio-

temporally sampled stimuli that the machine vision system will have to work with. However, it will be argued in chapter two that in general this difference will not substantially alter our approach to the problem at hand.

The first and most basic question that will have to be asked is what form the computation is going to take. There is a choice here since one could conceivably attempt direct computation of the motion parameters from the intensity function and its spatio temporal derivatives [2, 64]. This method has been proposed recently in restricted cases, like motion of planes or pure rotational motion. A generally applicable strategy according to this approach seems difficult and is yet to emerge. The other alternative approach, which is adopted here, is modeled after the abstraction hierarchy idea encountered previously. The two schemes are shown in figure 1.1.

As mentioned before, the chosen avenue for the investigations is motivated by the connectionist paradigm and the associated notion that the computation is structured so as to compute successive *invariant* levels characterized by a small set of parameters, as in the *parameter net* formalism of Ballard [9]. Such a methodological orientation dictates that the constraint relations between the parameters in adjacent layers be kept as simple as possible. In addition, it is desirable to minimize, as far as possible, the size of the parameter sets describing the invariants at each layer. (Later, this cardinality is referred to as the dimensionality of the parameter space corresponding to a particular invariance layer, a usage whose purpose will

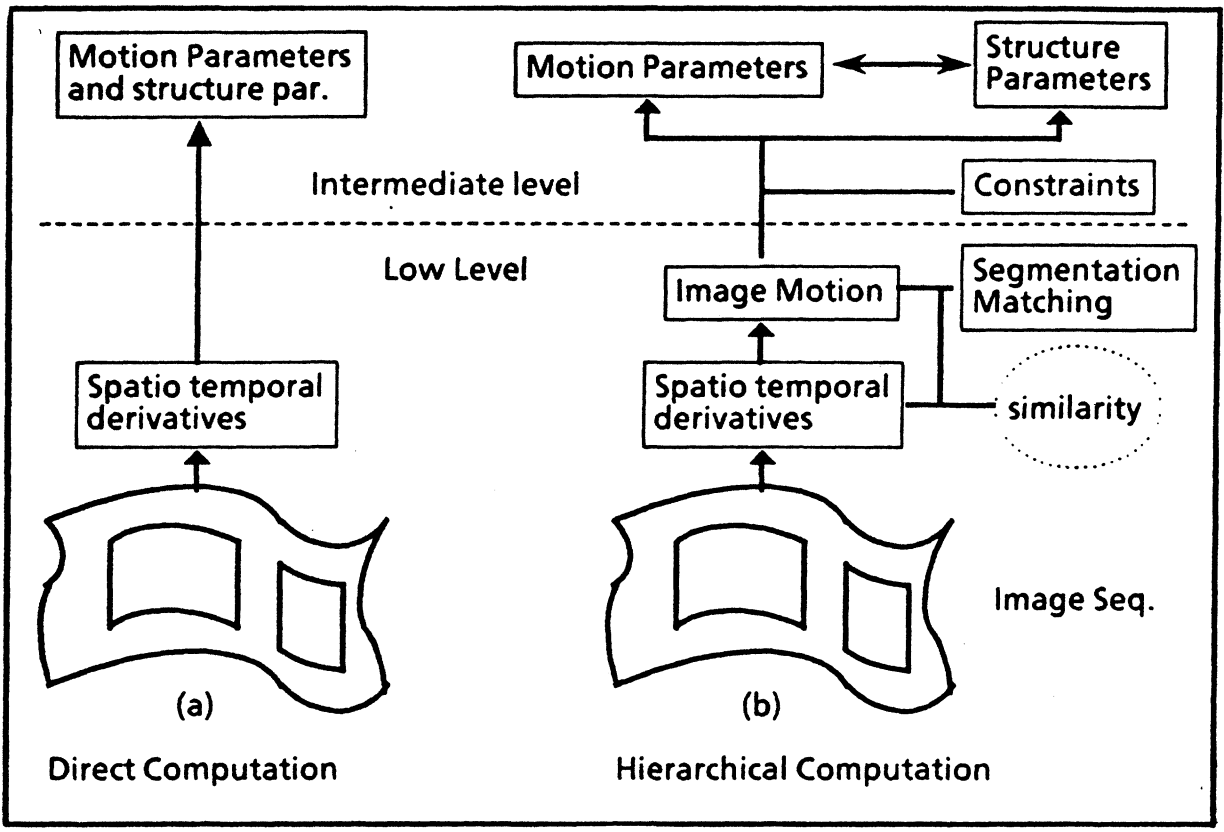


Figure 1.1 Alternative models for Motion Analysis

become clear when we introduce the notion of the hough transform as a general computational paradigm for parallel algorithms). The reason is that when we envisage, a highly parallel implementation of a computational scheme in the connectionist form proliferation of dimensions and complexity of constraints cause exponential growth in units and connections.

Now it is possible to answer the question as to whether all the layers in the proposed computer model are really necessary. Notice that in the direct computational model one is constrained to handle the motion and structure parameters together, necessitating higher dimensional parameter spaces and complex constraint relations for the parameter computation. On the other

hand, the layered model can be shown to deal with smaller parameter spaces and simpler constraints linking them (see chapter three). The demonstration of this fact was, in fact, a principal goal of this thesis.

The next section will paraphrase the subject matter of each of the chapters of this dissertation and indicate the contributions made. Some idea of the nature of the various layers, with reference to figure 1.1 will also be given.

1.4. Outline of the Dissertation

The structure of the proposed computer model for motion perception is given in figure 1.1. Chapter two deals with the computation of image motion. The computation is basically achieved in two stages, involving the computation of image features using local image filtering, followed by a cluster based matching and segmentation algorithm for estimating the image motion. The next chapter looks at the constraints governing the computation of the rigid motion parameters and structure. Investigation centers around determination of the nature of the computation, ways of segmenting the structure and motion parameter computation, and ways of resolving interpretation ambiguities when they arise. Chapter four details algorithms for motion perception from computed image motion. The constraints used, are derived based on the analysis of chapter three. The overall computational paradigm employed for this proposed algorithm is

called the *hough transform*, which is shown to be parallelizable and conceptually simple. Chapter five examines some of the difficulties faced by the final stage of the computation. An active tracking mechanism is shown to lead to considerable simplification of the computational requirement. Finally the last chapter concludes by reiterating the goals of the research and the results of the study.

The following subsections discuss some of the key ideas pertaining to the various following chapters of this thesis.

1.4.1. Image Motion Measurement

The goal here is to examine models for image motion, and determine how they can be computed. Intuitively, as well as from psychophysical evidence [36], it is seen that two dimensional velocity or optical flow is an adequate and useful representation for image motion. Optical flow captures the motion and structure information in the retinal image flux and is thus an abstraction useful for theoretical analysis. Schemes for optical flow measurement proposed in the literature are applicable only under restrictive circumstances. A common difficulty encountered occurs in image regions where contours are present. In this situation, components of the optical flow normal to the local contour orientation can be measured (Marr and Ullman [56] call this the aperture problem).

The retinal motion representation used in this research is a discrete form of optical flow. The measurement method is based on matching, "interest points" obtained by convolving the image frames with a set of feature masks and applying a simple decision rule for selecting or rejecting particular retinal locations. The relation between the discrete and the continuous representations (e.g. optical flow) is analogous to that between the chord and the tangent to a continuous curve at any location. The curve referred to is the interpolated trajectory of a retinally projected world point.

A problem with retinal motion measurement has been the difficulty faced by researchers in segmenting motion fields generated by more than one moving object [33]. The clustering approach adopted in our proposed model provides a uniform scheme for dealing with both the *matching* and the *segmentation* problem. Local interaction between motion vectors is modeled by a *similarity* function similar to one used in [69]. This approach is more flexible compared to using the mathematical notion of smoothness to constrain the motion field [44, 87], since the latter requires a dense sampling of the retinal space in order to estimate derivatives of the motion field. The method proposed here been tested on artificial as well as real data.

1.4.2. Constraints for Motion Analysis

No serious study of rigid motion perception can be successful without an understanding of the geometrical relationships that make it possible to compute three dimensional motion from retinally projected velocities (or displacements). The geometric analysis should be aimed at answering questions such as, what is computable and how simple are the computational steps required. One should be aware of the fact that the representations of the various entities to be computed at all stages of the computation must be chosen with care in order to ensure that they may be computed conveniently and there is no unnecessary redundancy.

In this respect, the choice of parameters to represent the motion of a rigid body has to be made. A simple solution is to represent the motion by the set of three dimensional velocity vectors corresponding to each observable point on the surface of the body. This is a redundant representation because, a rigid body, free to move in space has *six degrees of freedom*, therefore six parameters should be enough to describe its motion.

There can be many alternative forms of the six parameters, which are equivalent, but numerically not identical to each other. Examples of such representations are

(i) Translational and rotational velocity components of the body.

(ii) The instantaneous axis of rotation and the rotational velocity of the body.

The representation that is chosen, may be dependent upon the ease of computation and manipulation in the particular application domain. In most of the geometrical analysis given subsequently, the velocity representation for motion is used. The reason is that, this differential approach leads to simplicity in the algebraic relations used in the analysis without diluting the concepts underlying the mathematical characterization of the problem domain.

The geometrical transformation in the eye, or camera, giving rise to the two dimensional image from three dimensional scenes is called *perspective* or polar projection (refer to chapter three). Another model of transformation is the *orthographic* projection, which is an approximation of polar projection. The constraint equations obtained from the differential analysis embodies a "small" motion approximation. An understanding of the small displacement approximation is essential in order to determine under what conditions the constraint equations are valid and what are the errors introduced due to the quantization process that approximates differentials by differences. These issues are examined later.

It is known that [28, 84, 89] a single monocular observation of the optical flow field may not be enough to determine the three dimensional motion parameters uniquely. This ambiguity is seen, for example, in the

motion of a planar surface. Some of the algorithms that have been proposed for recovering motion parameters from discrete retinal displacements, have been analyzed to ascertain the conditions under which the computation leads to unique results. However, there has been no examination of the uniqueness question that is independent of any particular algorithm.

An analysis of the constraints which form the basis of any approach to motion perception leads to the following results:

(1) The motion ambiguity for planar surfaces can be resolved when the orientation of the plane is known, even partially, meaning tilt angle but not slant is available.

(2) In general there can be at most three interpretations of the optical flow field. Hence any local analysis, e.g. involving spatio temporal derivatives of. flow, must involve nonlinear equations (at least cubic), in the absence of shape information.

(3) If the three dimensional velocity of the rigid body under observation varies smoothly, then observation of the flow field at two or more time instants can determine the motion uniquely.

(4) Local shape information (surface orientation) is a powerful aid to motion perception.

The analytical results outlined above lead to an understanding of the theoretical basis of any motion perception algorithm. It also highlights the

fact that the task is difficult due to the inherent nonlinearity and the large size of the parameter set.

1.4.3. Algorithms for Motion Perception

Chapter four deals with algorithms for rigid motion perception, based on the analysis in the previous chapter. Some of the basic principles are demonstrated by computer simulations using synthetically generated data.

The computational principle underlying the design of the algorithms is the hough transform [7, 9, 18]. The idea derives from histograming in parameter space. Instances of the constraint hypersurfaces "vote" for parameter values that are compatible with it. The parameter estimated to be the most likely candidate, compatible with the global set of constraints, is the one receiving the largest number of votes. This vote counting can be implemented in parallel, by a connectionist network. However, as mentioned before, the number of units and connections grow exponentially with parameter space dimension.

The hough paradigm has been explored in the domain of motion parameter estimation. Some of the limitations of the approach, brought on by the nonlinearity of the constraint equation are examined and heuristics are suggested to overcome them.

1.4.4. Active Tracking Constraints

When a mobile system has the ability to visually track points in its environment, it can be shown that the mathematical relations that govern the determination of the motion parameters become considerably simpler. One might suppose that the demands of a tracking system might overwhelm its advantages. In other words, could the requirements of such a system be more difficult to achieve than the original problem of static motion measurement? It will be argued that this is not the case, since the tracking of the image of an environmental point is well within the reach of current technology, once that point has been identified.

The mathematical advantages of tracking: As we have seen, there are powerful advantages to designing a motion interpretation system based on tracking. The arguments in the foregoing sections have been mostly confined to the retinal structure in the flow field. An important point about the tracking regime is that it only needs retinal motion measurements for its sustenance. One can expect the matching of eye motion to the retinally projected motion of the imaged scene to facilitate the three dimensional motion measurements. This is indeed the case, in fact it will be seen that the following hold true:

(1) In the monocular case the number of parameters in the motion constraint equation reduces by one, without any increase in the degree of the nonlinearity.

(2) When the tracking is done by a system of two cameras whose relative positions and orientations are known, then the constraint equations reduce considerably in dimension, in addition to being linear in the parameters. In this case observation of the optical flow at just two points is enough to determine the motion parameters completely*

(3) For binocular viewing, it is necessary to combine the optical flow fields from the two eyes. However, this is not necessary, when the observation period extends to more than just one instant of time. In this case one can obtain *closed form solutions* for the rigid motion parameters without the necessity for binocular fusion.

1.4.5. Summary

This research is concerned with the computation of rigid motion parameters from spatio-temporally varying retinal stimulus. The problem is approached in three stages. These relate to the *mathematical and geometrical relationships* that exist between the three dimensional parameters and their retinal counterparts, the *representations that can be computed at various levels of the computational process* to facilitate the perception process and the *structure of the computational processes* themselves.

Chapter Two

Computation of Image Motion

2.1. Introduction

In keeping with the hierarchical model for the interpretation of visual motion, the first task that is investigated concerns the measurement of the *image motion stimuli*. To talk meaningfully about this latter measurement process it will be necessary to define the input and output quantities and various intermediate representations. The objective is to study the problem from the point of view of machine vision. However, in many cases the approach adopted is based on, what is believed to be, certain principal aspects of biological vision.

Mathematically, the input is a three dimensional (spatio-temporally varying) intensity function. The spatial coordinates (x,y) of this function $f(x,y,t)$ refer to the cartesian indexing of the retina or image plane. In reality however, as far as the human eye is concerned, the available input is a spatially sampled and temporally averaged version $\hat{f}(x,y,t)$ of the underlying function $f(x,y,t)$. Similarly, for the machine vision case, the

input is again a spatio temporally sampled version of the "real" image.

The temporal discreteness of the latter imaging situation has led visual psychologists to designate this type of visual input as *apparent motion stimuli*. The fundamental distinction between the "real" and apparent motion is stimulus continuity. However this distinction will not affect the proposed computational algorithms under the proviso that the spatio temporal variations (frequencies) in the underlying "real" image distribution $f(x,y,t)$ are not lost in the sampling process, and there is no aliasing. This will be called the *adequate sampling assumption*.

The second issue has to do with the determination of what to compute. In other words, what is an adequate explicit representation for image motion. An answer is provided by the notion of *optical flow*, a concept attributed to J.J. Gibson [36]. The optical flow field can be thought of as the retinal projection of the three dimensional velocity field that could be thought as the representation that describes the motion of rigid objects and surfaces. Of course, as will be seen in chapter three, for rigid bodies, there is a much more parsimonious description of the motion, than the three dimensional velocity field. None the less, it will be assumed that the optical flow representation will serve as an adequate representation for image motion [63].

Optical flow is an idealistic notion, and to measure it requires a continuous motion stimulus, which neither the biological nor the machine

vision systems have available to them. But, as mentioned before, by the adequate sampling hypothesis, it is assumed that the sampling process does not entail any loss of information. So it will be claimed that optical flow can, in principle, be recovered. There are essentially two alternative stages of processing where the transition from discrete to continuous may be made. Correspondingly, there are two distinctive styles or classes of image motion measurement algorithms:

I. **Continuous Techniques:** The sampled image function $\hat{I}(x,y,t)$ is interpolated at the onset of the measurement process, to obtain the "real" image $I(x,y,t)$. The subsequent processing can then be based on continuous rather than discrete transformations and operations.

II. **Discrete Techniques:** The discrete point to point displacements are computed over the quantized space time dimensions. The perceptual system then performs smooth interpolations over a small "integration" time period when the spatio temporal trajectories of the observed "tokens" are smoothed to obtain a sparse sampling of the optical flow field at the retinal locations where the match tokens were found.

The retinal motion measurement algorithms proposed so far belong to either of the above classes [61]. Unfortunately, all such algorithms, be they continuous or discrete, suffer from ambiguity or local indeterminacy. For instance in the discrete case locally there may be more than one possibility for finding tokens to match a particular token item. Marr and Ullman call

this the aperture problem [56]. The problem arises due to the local nature of the measurement algorithm and hence a limited field of "attention" or aperture. Thus if the aperture permits the viewing of only a part of some smooth contour (e.g. a straight line), then due to the fact that there is no distinguishable token on the contour, the motion of points on the contour can be constrained (figure 2.1) but not determined exactly.

Thus if the instantaneous optical flow field is denoted by $\dot{\mathbf{r}} = \{u(x,y), v(x,y)\}$, and the constraint available in the local aperture is $C(f(x,y,t),u,v) = 0$, then at every sampled image location (x,y) there are two unknowns to determine (i.e. the values that the functions u and v take), but only one equation.

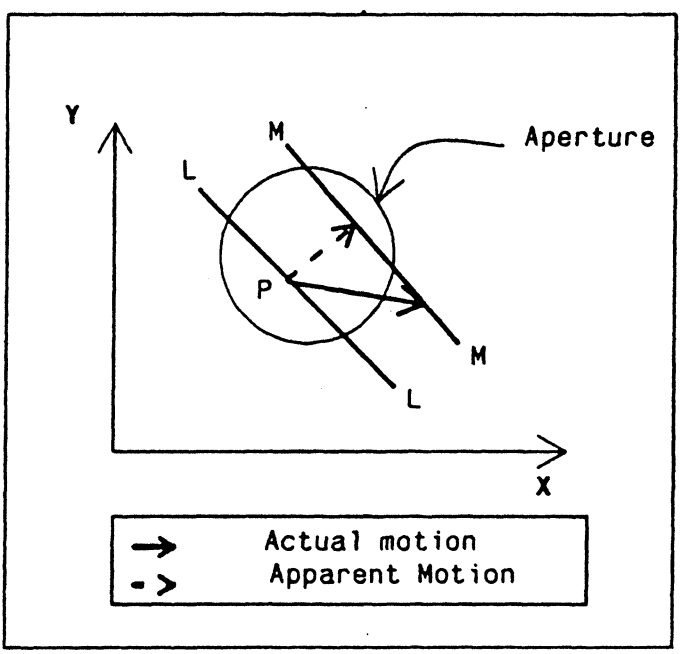


Figure 2.1 The Aperture Problem

To overcome the local ambiguity additional constraints are necessary. The main assumption used by other researchers is that the velocity field is globally smooth. In terms of the variational calculus, one assumes that the constraint equation is not exactly satisfied, but $C(\cdot) = c$ where e is the amount of error. Now if the smoothness criterion is given by the functional form $S(ujv) = 0$, then the measurement problem can be posed as minimizing $\wedge = (?(\bullet) + X5(*)$ with respect to 'u' and V, where X is aLagrange multiplier.

The above formulation has been the hallmark of a number of continuous optical flow measurement algorithms [40, 44, 87], One of the primary problems with the above class of, spatio-temporal gradient based, methods is that the convergence criteria and rate for the relaxation process are not known. This lack of performance bounds limits the applicability of such methods.

The second type of continuous formulation seeks to eliminate the inherently iterative/search nature of gradient based algorithms. There are two ways in which this has been attempted, one is to implement digital filters that are sensitive to time varying intensity patterns, and which purportedly mimic the spatio temporal receptive fields of the biological system (an example can be found in [34]). However the problem here is that it is hard to determine exactly what such filters indicate quantitatively, although qualitatively the outputs of such filters may prove to be useful for discrimination and segmentation purposes. Another type of motion

computation involves assumptions regarding local structure of the moving surfaces. For instance assuming that the visible surfaces in motion are locally planar, leads to the locally second order flow fields which may be easier to measure [1, 90].

Surprisingly, the class of discrete algorithms, has not been explored with the same vigor applied to the study of the continuous methods. The paradigm of token matching is the dominant strain for such methods. Contrary to continuous methods the main operating criteria here are:

(i) The motion measured be due to the geometrical projection moving features in three dimensions.

(ii) The measurement be immune to variations in lighting and viewpoint.

(iii) No elaborate form analysis precede the actual measurement operation.

The usual approach here [14, 85] is to assign different a priori probabilities or confidence to the competing match vectors and to chose the best set of non conflicting matches based on some global compatibility measure. Of these two methods, the first [14] employs a solution method that is ad hoc. The notion of similarity is never exactly quantified. The claim is that their confidence update rule captures the notion of local similarity, however their method never makes it clear the exact nature of this interaction. The rate of convergence of the algorithm is also not shown.

On the other hand Ullman's exposition [85], which also concerns discrete matching, seeks to develop a mathematical theory of visual motion computation. While this *minimal mapping theory* is in itself, an exemplary work in the field of scientific exegesis, it makes certain strong assumptions and leaves certain questions answered. The main idea is simple and elegant, and proposes the choice of matches to minimize the entropy of the global field of matches. Each velocity v is associated with an entropy measure $q(_v)$ =-logp(t;), where p() is the probability that v is the true velocity. Thus the idea is to assume q(v) as the cost of assuming velocity v, so that it can be minimized globally. In other words he is looking for the maximum likelyhood solution.

The problems with this approach mainly stem from the fact that to translate the above idea into a working mechanism, one has to make some simplifications. The simplifications that Ullman proposes require the assumptions that the probability distributions of retinal velocities that are obtained at different image locations, are *independent* and that the probabilities are inversely proportional to the velocity magnitudes. It is very hard to justify the first assumption, while the second is a very coarse approximation which is not entirely justified by empirical data presented in [85], The remaining objection to the method is its computational complexity. Although the algorithm is formulated as a linear programming problem, the gradient method proposed for its solution need not converge

to the desired solution, and in fact could lead to cyclic behavior under certain extreme conditions. Hence it is difficult to envisage a network of parallel computing elements implementing this algorithm, within the limits of the performance constraints imposed by a "realistic" (or biological) computational devices.

The discrete algorithms described so far, we will classify as *discrete/iterative,* since, in their proposed forms, they involve local search with ill understood rates of convergence.

In contrast we believe that a computational theory of image motion should try to satisfy some desirable properties not encompassed by the above motion measurement algorithms. For instance:

(i) The velocity estimates for a single region should^collected together, while those for different regions should separated out. Thus not only should our theory show how to handle local similarity but also recognize boundaries where dissimilarities occur.

(ii) The formulation of the algorithm should also bring out the complexity of the computations, and justify the simplifications introduced to reduce the complexity.

(III) Since this is basically a low level visual computation algorithm, the style of implementation should be a major concern during theory formulation. This is in some sense in opposition to the well known

position by Marr [57] that *theory, representation* and *implementation* are independent of each other. While this may be so for high level symbolic computations, it is evident that at lower levels this epistemic neatness is infeasible. For instance if we agree that a highly parallel hardware is desirable and even necessary, then in order to realize performance/cost criteria like dynamic range, speed and efficient encoding, we might have to resort to well established devices like "coarse coding", quantization of the measured parameters at various degrees of coarseness, and so on. It is hard to imagine a probabilistic global optimization process being implemented under such conditions. This is essentially the connectionist argument [6, 31] applied to a concrete case.

The image motion measurement method proposed here, is based on the hough transform in a more general form than is normally prevalent in the vision literature [7]. The method is clustering based, where the complexity of the measurement process and the segmentation process are treated uniformly. The complexity of the general problem is very clear in this approach and the simplifications that allow us to obtain biologically plausible parallel implementations using minimal spanning tree algorithms seem justified based on simulation experiments. Because of this tighter complexity and computational bounds, we classify this method in a class distinct from the other discrete methods. The latter class is the

discrete/non-iterative category of algorithms for image motion measurement.

The classification of the algorithms described above is summarized in Table 2.1.

In general, the methods in categories I,II and III are specialized and work reasonably well only in restricted domains. Examples of restrictions imposed are: uniform illumination, smoothly varying reflectance, being able to locate smooth zero crossings (of the Laplacian of the image intensity) contours and local planarity of the moving surface. One shortcoming of these approaches is that they deal primarily with movement of a single object in the environment or motion of the observer. Some of the above techniques are sensitive to noise. The proposed cluster based approach seeks to remedy these lacunae.

	Continuous	Discrete
Iterative	I. Spatio-temporal gradient methods [33, 38, 40, 44, 56, 62, 65].	II. Matches compatible with local constraints [14, 66, 85]
Non-Iter.	III. FIR filters or local polynomial approximations [34, 90]	IV. Proposed cluster based approach

Table 2.1 Classification of Image Motion Measurement Schemes

2.2. Overview of the Clustering Algorithm

The algorithm, can be implemented very easily by a parallel network of relatively simple computing elements and is motivated by the connectionist paradigm in AI and cognitive modeling [31, 32]. The structure of the algorithm closely models the parameter network formalism of [9]. The measurement of image motion is performed in the lower levels of a hierarchy of computing layers of neuronal computing elements (see figure 2.2).

The layered structure reflects an organizing principle: that vision can be viewed as computing key parameters at different levels of abstraction.

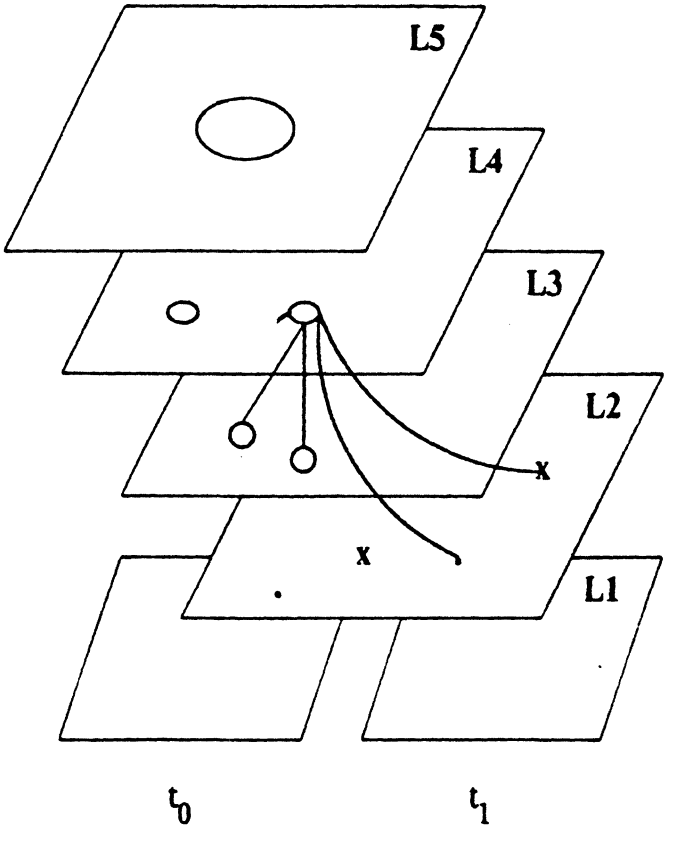


Figure 2.2 The Motion Perception Hierarchy

Furthermore, a natural progression of abstraction layers start from low levels (i.e. the image intensity function) and evolve to high levels (object tokens). At each level an important concept is the size of the local spatial domain in the image, over which the parameters at a level can be modeled as invariant. In analogy with the use of the concept in biology, we term this domain, the *spatial receptive field (SRF) of a parameter.*

The idea of a parameter's SRF can be best understood by an example. Figure 2.2 illustrates the hierarchical layered structure in the computational model of motion analysis under discussion. The lowest (i.e. the rawest) level of representation is depicted by the plane, L1, consisting of the image intensity function. In the first stage of processing locations of significant intensity change in the image are computed (layer L2) - note that change units have a very small SRF. The next layer L3, computes the retinal motion parameter (e.g. optical flow), indexed by the image frame positional coordinates (x,y). The SRF of flow parameter value is much greater than that of the change units. The parameters computed in the following layer L4, can be thought of as motion vectors that are not specific to particular spatial locations, but indicate the distribution of velocities in a "window" of the image. Later descriptions detail how clustering in this "space" of location independent (relatively speaking) motion vectors, helps to establish correct matches in the token matching layer L3. The next higher layer computes the parameters of motion of the imaged surface (over the region,

which need not be known to the measurement process a priori, corresponding to a single moving body).

The flow of data is from the lower layers of this hierarchy to the upper layers. There are exceptions, however, since the clustering layer L4 influences the matching process in the lower layer L3 of velocities (therefore the structure is not a strict hierarchy). The design of the computation in the layers L2,L3 and L4 and how retinal motion is computed cooperatively is described subsequently.

The scheme of computation of the retinal motion has the following steps :

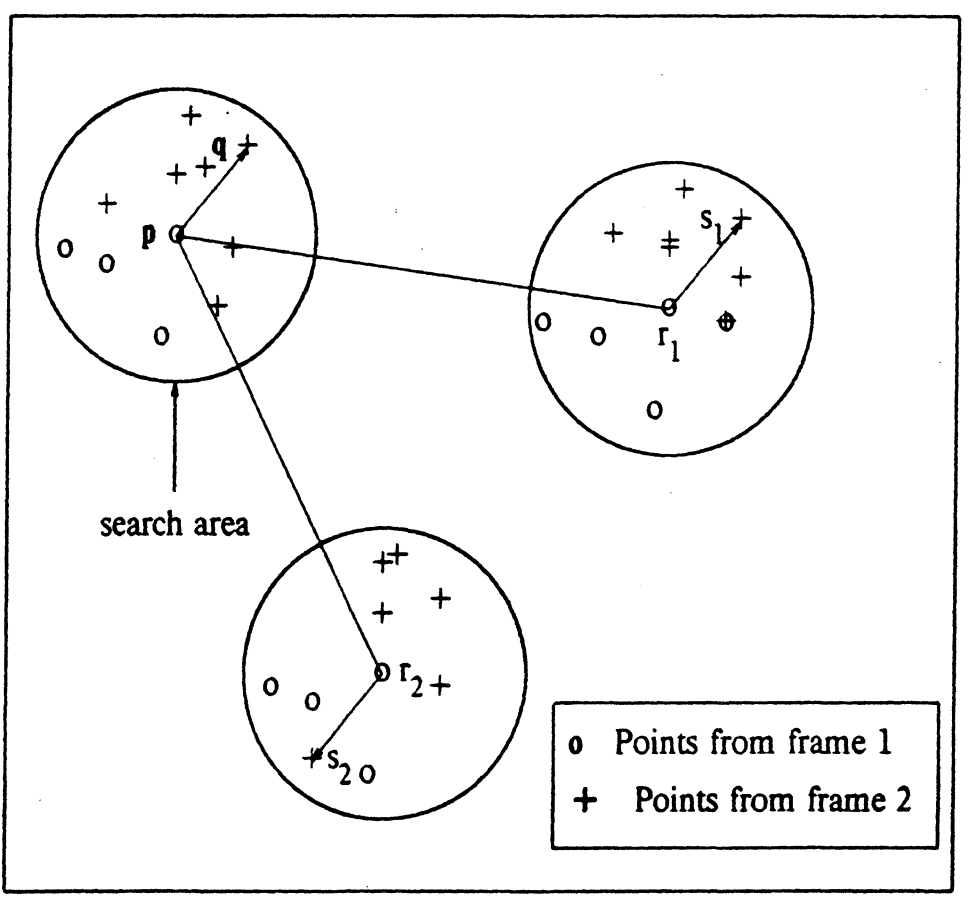


Figure 2.3 Matching Interest points from two frames

(1) *Location of points in the image with significant intensity variation (contrast).* The goal is to select locations in the image where there is significant (with respect to its neighborhood) contrast, yet there is no orientation specificity. This means that two criteria must be satisfied:

(a) The contrast variation at and around the selected location should be high. This could be measured by means of a center surround operator like the Laplacian or DOG operator [55, 57]

(b) A good edge operator should not respond to intensity distribution at the same location, or have multiple weak responses.

The interest operator that was used is described in [12]. This operator decomposes the local intensity distribution around a point in the image into a set of basis functions. The selection of the point then depends on the relative responses of the "edge" and "extremum" subspaces in the basis set. A following section will detail the design and operating characteristics of this method.

(2) *Measurement of image motion** It is assumed that the velocity field is locally similar, except for a small number of places where motion boundaries occur. Each point selected in a given frame (i.e. time instant), can potentially match another point in its neighborhood, selected from a later frame. The process is depicted in figure 2.3 where the large circles indicate the areas searched to obtain plausible matches.

To determine the goodness of the match, each of these "velocity units" (or plausible match vectors) evaluate the support it receives from nearby velocity units. This scheme was used with great success for stereo matching by Prazdny [69]. Finally only those matches that can muster more support than competing matches get selected.

(3) Clustering in image motion space. As the velocity units are evaluating their support, they also "vote" for non location specific velocity units in level L4. The units in level L4 then cluster around similar (the Euclidean distance metric is used) units and support each other. This helps to remove the outliers among them. The units that belong to some cluster are then retained and the rest are deleted. The surviving units then mediate the matching process in the lower level L3. The basic idea is illustrated in figure 2.4.

The discussion so far would seem to imply the requirement of two "snapshots" of a dynamic scene, taken at consecutive time instants. This is not a critical aspect of the method, being only used for ease of explanation. In fact it adapts very easily to a sequence of temporal frames of a changing scene. In this case all we need to do is to introduce a temporal decay rate for the accumulated support for the velocity units in layers L3 and L4. The algorithm has been tested with synthetic images comprising spheres and planes, which where painted with random dot patterns. This was mainly done so that the computed motion vectors could be compared with the

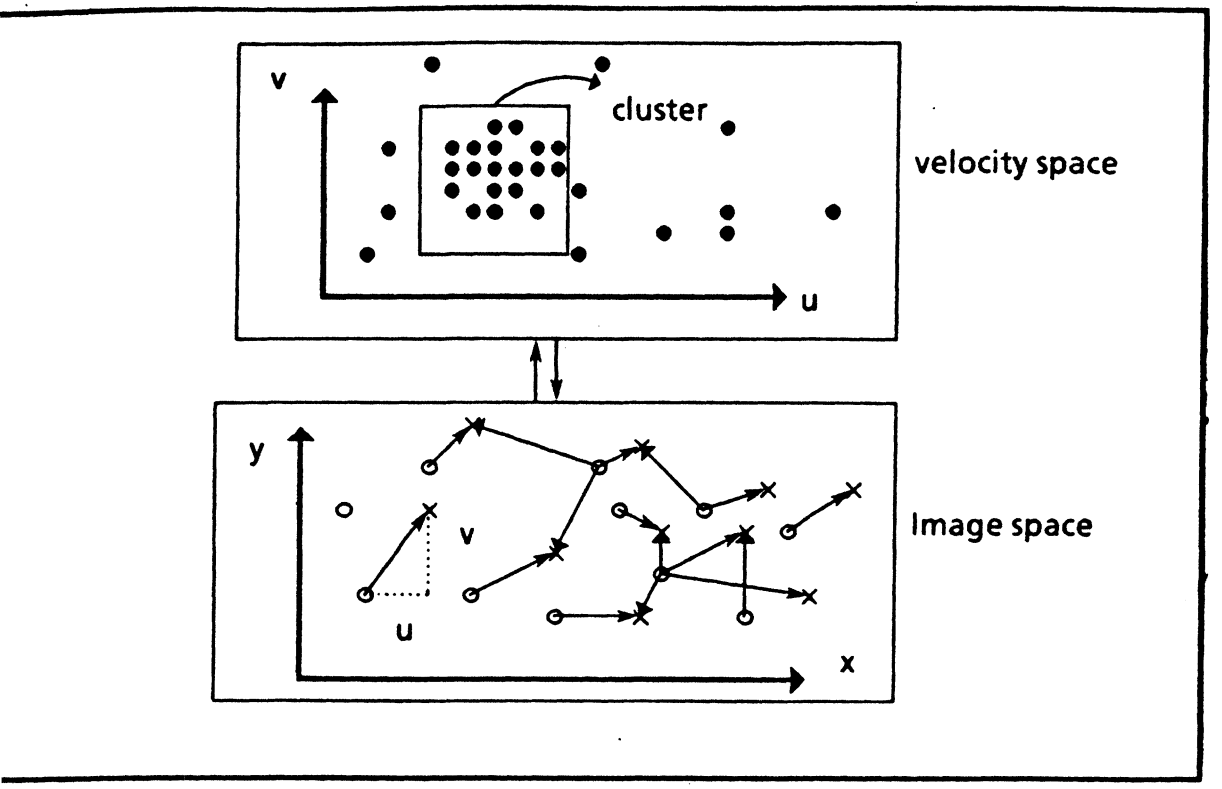


Figure 2.4 Clustering in image velocity space

actual values. The experiments with synthetic data show that multiple moving bodies and as much as 20% random noise points can be handled by the algorithm. The following sections detail the various parts of the algorithm and the experimental results.

2.3. Computing Interest Points

An interest point is a point in the image (actually a small neighborhood) which has *properties* that distinguish it from its neighboring points. The properties in question may be simple, like gray levels, or sophisticated ones indicative of the local topography of the imaged surface. Previous approaches to finding interest points are exemplified in the work of

Moravec, Kitchen & Rosenfeld, Nagel, Davis, Sun & Wu, Fang & Huang ([24, 27, 51, 58, 62]).

The difficulty of locating interest points for matching stems from the fact that it is difficult to specify exactly what should be the desirable characteristics of such feature points. On the other hand it seems clear that the following properties are in general desirable:

(i) The detection and localization of these points should be fairly straightforward. In other words, the features that trigger the detection of such points must be computable by examining a fairly small support region in an image.

(ii) These points should be preferably be sparsely distributed in the image. A measure of such sparseness depends upon the support region size for the subsequent matching algorithm. Thus one performance parameter could be a measure of the average number of false matches that have to be considered for every correct pairing of points from two frames. It has been reported that human performance degrades significantly when this number increases beyond four or five [78].

(iii) The feature that characterizes the points must be stable. This means that small changes to viewpoint and illumination should not affect their determination.

Concrete proposals for specifying and locating feature points for matching, fall into the following classes:

1. *Grayvalue corner Selection:* The feature is restricted to corners in the image intensity "terrain". The method of detection usually consists of finding the extrema of the spatial gradient of the intensity function [51, 55, 62].

2. *"Interest point" selection:* The idea here is to pin down image patches, where the intensity variation profiles are distinctive in the support region. This distinction can be mathematically specified, for instance, by measuring the variance in the pixel intensity values and selecting locations where it is maximized within a local support region [58]. Another method would be to chose patches with sharp autocorrelation functions [16].

3. *Selection by "topological analysis":* This method attempts to label the intensity terrain with labels such as hill (maxima), pit (minima), ridge/ravine (line), saddle, table edge (edge) and flats. The idea is that these labels being relatively viewpoint and illumination independent, compared to raw intensities and gradients, can be used to perform sophisticated correlation type of matching algorithms [38].

The problem with corner detectors and the topological analyzers is that the support region needed is quite large. This is because of the fact that it is necessary to obtain some locally smooth approximation for intensity

distribution in the support window, in order to be able to compute the necessary gradients. This requirement of higher order derivatives of the intensity function (e.g. the hessian) and the attendant computational complexity diminishes the attractiveness of the above methods.

The variance based interest operator is poor at contour suppression (especially for edges oriented at small angles with respect to the horizontal or the vertical directions). Furthermore, being intensity based, it is sensitive to viewpoint and illumination changes. Finally, the idea of operator design based on maximizing autocorrelation is yet to be translated into a successful design.

The method of interset point selection suggested subsequently seeks to combine some of the positive aspects of the above mentioned techniques. The salient advantages of this operator are:

(a) The selection process is essentially linear (convolution) and amenable to parallel implementation.

(b) The features that are key to the detection process are not intensity based. In fact they have the flavor of the labels computed by the topological classification methods, *without* the attendant computational complexity.

(c) Contours are suppressed at all orientations.

(d) The method does not depend upon thresholds that have to be determined a priori.

The principle underlying the proposal is based upon the comparison of the outputs of isotropic feature masks. The basic scheme involves comparing the responses of some edge operator with a center/surround operator like the laplacian. This scheme is depicted in figure 2.5.a, where the edge response is provided by the Sobel operator. The normalized values of the edge response (solid curve) are compared with the response of the laplacian (dotted curve) are shown in figure 2.5.b, for a step edge profile. The responses are plotted as the operators are applied along a straight line path

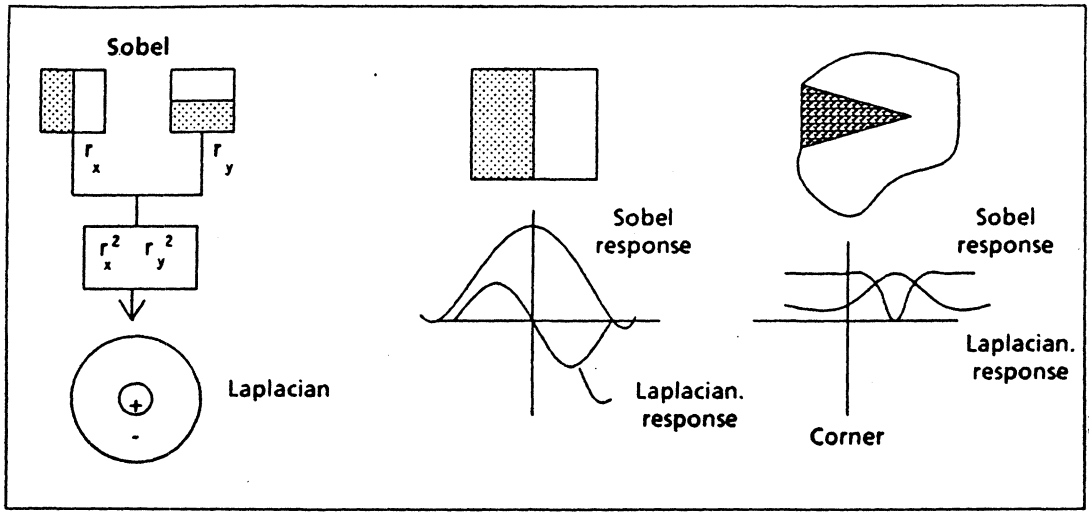


Figure 2.5 Feature Selection by Comparison of Operator Outputs

perpendicular to the edge. So we see that if we were to compare the operator response at or near the step edge, the edge response will always be greater. However the situation is different when the image profile is shaped like a corner. In the latter case (figure 2.5.c), the response of the laplacian is stronger than that of the sobel operator.

It seems that this simple scheme should be capable of selecting interest points in an image. However, there are some shortcomings of this approach:

(1) The sobel operator is a poor detector of edges oriented away from the vertical, horizontal or the two diagonal directions. Hence the above scheme may select points along such edges.

(2) There are many distinctive variation patterns of the image intensity function that cannot be detected by such a scheme (see figure 2.6 and table 2.2).

The above problems make it necessary to reexamine the operator design if we want to make the relative response measurement criterion work. To alleviate the first problem one could resort to directional edge detectors, like the Canny operator [20]. The difficulty of such an approach is that the scheme entails the the replication of the directed operator at some intervals of orientation angle. Furthermore, since the above scheme is based upon comparison of operator responses, it is easier to break up the image vector space (this concept is explained subsequently) into orthogonal bases rather

than resort to different types of operators for different types of features and then calibrating the relative responses on a large number of sample images.

As a solution to the first problem mentioned above, the sobel masks were augmented with two more 3x3 masks. This arrangement is reported to be more isotropic in its response to edges of arbitrary orientations [35].

To minimize the second problem, it was decided to investigate the response characteristics of rotation invariant filter masks [23]. These can be represented mathematically by

$$M(r,\phi) = h(r)e^{jn\phi} \qquad n = 0,1,2,3,...$$

where $j=\sqrt{-1}$, $h(r)$ is a radial weighting function and (r,ϕ) are polar coordinates for position. It is generally preferable to take $h(r)$ to be a gaussian function. In our case, since for simplicity the masks were limited to 3x3 size, $h(r)$ was taken to be constant.

In this case the operators of various orders (corresponding to values of 'n') can be interpreted as follows:

(a) n = 0 : Averaging operator.

(b) n = 1 : Step edge operator.

(c) n = 2 : Line Detector (second and third operators in figure 2.7)

For reasons of economy, the operator size was confined to 3x3. However, these operators were applied at various image resolutions by smoothing and sampling the intensity distribution but keeping the operator size fixed. A

very important property of the set of mask described so far is that with the addition of a saddle mask, we have a set of nine 3x3 masks which form a complete orthogonal basis for the space spanned by the nine dimensional vectors defined over the real field (figure 2,7).

In addition, the space spanned by the feature basis set can be thought to be subdivided into three components which we term the *edge, extremum* and *average* subspaces. The addition of the line and saddle mask (the fourth operator in the extremum set in figure 2.7) proves to be more powerful for detecting intensity profiles that show a high degree of curvature, but are not like step edges. Thus for example, acute corners have the characteristics of line terminations and hence trigger the extremum detectors due to the presence of the line operators in the latter space.

2.3.1. Feature Classification fay Orthogonal Decomposition

The image *f{x,y,t)* is a three dimensional function. However, we concern ourselves with a time slice of this function at time =<* thus obtaining a two dimensional function

$$\mathbf{J(*,y)} \quad = f(x_9 y, t_{\%})$$

An image vector at a location *(x,y)* is formed by concatenating the **rows** of the following 3 x 3 image patch

The the image vector *T* belongs to a 9 dimensional *Vector Space* defined over the Real field.

I(x-1,y+1)	I(x,y+1)	I(x+1,y+1)
I(x-1,y)	I(x,y)	I(x+1,y)
I(x-1,y-1)	I(x,y-1)	I(x+1,y-1)

$$\vec{i} = [I_1, I_2, I_3, \cdots, I_9]$$

where I_1 is $I(x-1,y-1)$, I_2 is $I(x,y-1)$ and so on, alternatively

$$\vec{i} = \sum_{k=0}^{8} I_k \cdot \mathbf{e}_k$$

where $\mathbf{e}_k$ is the k^{th} column of the 9x9 identity matrix.

The image vector as defined previously, is represented with respect to the basis $\{\mathbf{e}_i\}$. The components, therefore, by themselves do not convey any information regarding the local topography of the image.

When we define the image in this manner, the operation of convolving the image with a given point spread function or correlating with a particular feature template can be expressed with respect to the vector inner product. Thus convolution becomes

$$I * h = \sum_x \sum_y \vec{i} \cdot \vec{h}$$

where $\vec{h}$ is the vector representation of the point spread function and $\vec{i}$ is the image vector at a point.

With the above interpretation in mind we freely interchange the terms *function* and *vector* in subsequent text. More importantly, thinking of point spread functions as vectors, allows us to transform the image vector into different *finite* basis space corresponding to the prototypical features that we are interested in. This transformation is wrought by a non singular matrix T

whose columns are the feature basis vectors $\{f_i\}$ Thus the image vector $\vec{i}$ is transformed into the vector $\vec{v}$ where

$$\vec{i} = T\vec{v} = \sum_j v_j f_j \tag{2.1}$$

The purpose of this transformation, in our case is to obtain a image code whose components correspond to the degree of match between the image function and the feature functions. In general, to compute the transformed vector $\vec{v}$ from $\vec{i}$ requires the solution of simultaneous linear equations. However, computation of the k^{th} component of $\vec{v}$ becomes simple when f_k is orthogonal to the other basis vectors in the set $\{f_i\}$. In this case, we have from equation (1)

$$\vec{i} \cdot f_k = \sum_j v_j f_j \cdot f_k$$

$$= v_k \|f_k\|^2$$

In particular, if the basis vectors are chosen so that they form an orthonormal set then

$$\vec{i} \cdot f_k = v_k$$

Since the image vector is finite dimensional we can design a orthogonal basis set for the space of the image vector. In addition, this basis set is constructed in such a way that the each basis corresponds to a feature primitive. Decomposing the image vector in terms of the new basis would give us a new set of components (or weights) indicating the *strength* of each of the features represented by the respective basis vector. This idea is

originally due to Frei and Chen [35]. Their purpose was to develop an edge detector which would not require thresholding after the convolution step. They convolved a 3x3 image region with nine orthogonal masks and compared the outputs of the edge masks with a set of "line" detection masks. The present method for interest point selection follows, in some sense, a strategy that is a dual of Frei and Chen's approach with a different set of basis functions.

As stated before, the set of basis functions in our model is built around feature primitives like edge, maxima/minima and saddle type variation. Since the image vector is nine dimensional (i.e. the operator size is 3x3) there are nine elements in the feature basis space. The feature space is divided into three subspaces:

1. The Extremum subspace defined by the laplacian, line and saddle masks.
2. The edge subspace.
3. The average subspace.

The basis functions used to define these subspaces are shown in figure 2.6. A key characteristic of the feature subspaces is the directional isotropy of their response patterns.

To test the applicability of the above image decomposition scheme a set of image profiles were devised to verify the ability of the interest operator to

1	2	1
0	0	0
-1	-2	-1

1	0	-1
2	0	-2
1	0	-1

-2	1	0
1	0	-1
0	-1	2

0	-1	2
1	0	-1
-2	1	0

The Edge Basis Functions.

-1	-1	-1
-1	8	-1
-1	-1	-1

0	1	0
-1	0	-1
0	1	0

-1	0	1
0	0	0
1	0	-1

1	-1	1
-1	0	-1
1	-1	1

The Extremum Basis Functions.

1	1	1
1	1	1
1	1	1

The Averaging function.

Figure 2.6 Interest Operator Masks

classify some idealized markings. The test images were 3x3 masks shown in figure 2.7. The results of the test are summarized in Table 2.2. The figures in the columns indicate the normalized responses of the operators to the respective image masks. The last two columns indicate the result of applying two different decision rules for interset point selection. It is seen that the simple rule which compares the outputs of the laplacian and the sobel operator is inadequate, although it is enough to discriminate between interesting and "edgy" regions in many instances.

No.	Av.(M)	Lap.(L)	Sobel(E)	Point(PS)	Edgc(ES)	PS> ES	L> E
1	0.34	0.95	0.00	0.95	0.00	yes	yes
2	0.34	0.12	0.58	0.98	1.16	no	no
3	3.01	0.00	2.31	0.00	3.47	no	no
4	3.01	0.00	2.45	0.00	3.47	no	no
5	1.01	0.36	1.16	1.21	1.74	no	no
6	3.01	0.00	0.00	2.01	0.00	yes	no
7	3.01	0.00	0.01	2.01	0.01	yes	no
8	6.34	0.95	0.01	3.78	0.01	yes	yes
9	4.34	1.65	0.01	5.36	0.01	yes	yes
10	6.67	0.83	2.05	3.68	3.47	yes	no
11	0.67	0.83	0.58	1.68	1.16	yes	yes
12	5.67	3.30	5.20	6.30	5.20	yes	no

Table 2.2 Operator responses to the test masks

2.4. Algorithms for retinal motion measurement

The *correspondence problem* is almost universally regarded as difficult. As mentioned earlier, it arises in the measurement of temporal image *disparities.* The problem is magnified for motion measurement, since the disparity in this case is not constrained, as in the case of stereo, to lye on a known line (*epipolar)* in image space. The overall scheme of thing is simple: select interest points in image frames and then decide which point from one frame matches another point from the other frame. If it is possible to obtain interest points that are sparse then correspondence is not difficult. Here sparseness means that the average disparity value is smaller than the average spatial distance between points in the same image frame. An interesting quantification for the degree of sparseness is due to Stevens [78] and is the number of false matches possible, on average, for a given match

neighborhood size.

2.4.1. The Matching Algorithm using local support

The algorithm proceeds from the interest point stage by forming all possible matches subject to a maximum limit on the magnitude of the match vector. This is equivalent to saying that the match neighborhood size is determined a priori. Each match vector then proceeds to accumulate

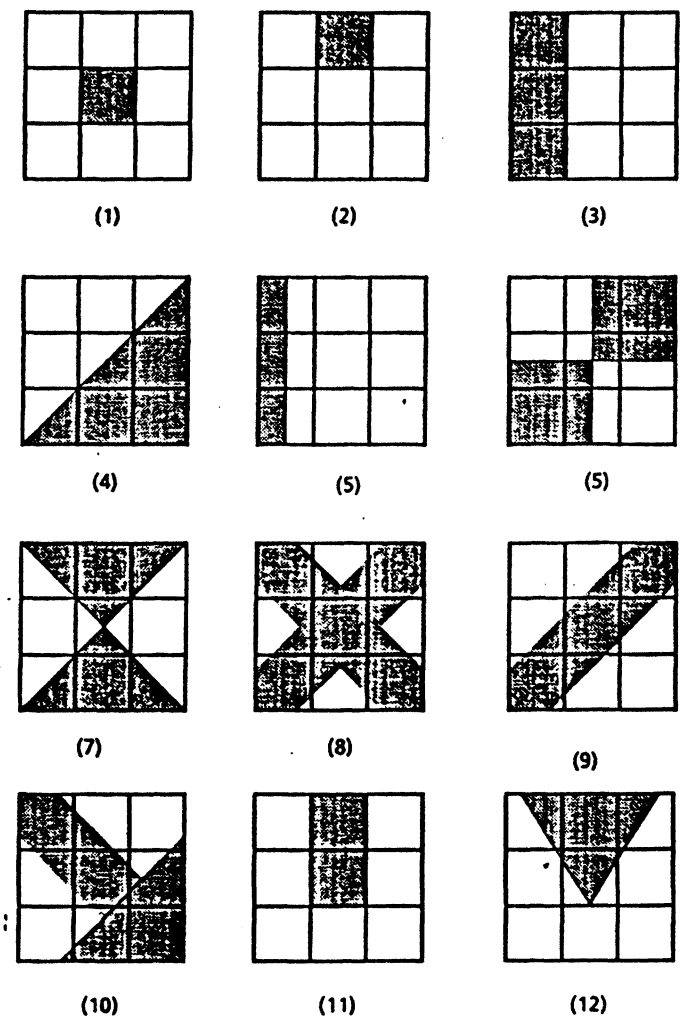


Figure 2.7 Image masks to test Interest Operator

evidence supporting its existence within a support neighborhood, which is larger than the match neighborhood. This scheme is based upon the assumption that the imaged surface depth varies smoothly (a similar scheme is reported to be successful with the stereopsis problem [69]).

To justify the notion of local support, consider optical flow (u,v) generated by a translating object. In this special case the constraint equations are

$$u = \frac{U - xW}{Z}$$
$$v = \frac{V - yW}{Z}$$

where (U,V,W) is the translational velocity in three space, Z is the depth of the object corresponding to the retinal location (x,y).

If the depth function $Z(x,y)$ is smooth then, to a first order approximation, the spatial rate of change in the optical flow is proportional to the spatial rate of change in depth. For instance consider the optical flow value at $p = (x_0,y_0)$. Let it be (u_0,v_0), also let the depth at p be Z_0. Then the difference between optical flow at p and a neighboring point $r = (x,y)$ is :

$$\delta u = u(x,y) - u_0 \simeq \frac{W}{Z_0^2}\Delta x + \delta Z\left(\frac{U - x_0 W}{Z_0^2}\right)$$

where $\Delta x = x - x_0$ and higher order terms involving $\frac{\Delta Z}{Z_0}$ and $\frac{\Delta x}{Z_0}$ are neglected. This leads to:

$$|\delta u| \leq X_1|\delta Z| + X_2|\delta x|$$
$$|\delta v| \leq \lambda_3|\delta Z| + \lambda_2|\delta y|$$

where X_1,X_2 and X_3 are constants for a local image neighborhood. Combining the above we have:

$$\frac{dst(u,v)}{dst(z,y)} < \lambda \frac{|SZ|}{dst(x,y)} + A_2$$

where $X = X_X + X_s$, $dst(u,v) = |\delta u| + |\delta v|$, and $dst(z,y) = |\delta x| + |\delta y|$. Thus the situation that arises here, is that the smoothness in the depth function relates to the smoothness of the displacement field The support that two candidate vectors (u_1,v_1) and (u_2,v_2) at retinal locations (x_1,y_1) and (z_2,y_2) respectively provide each other is given by the function $S[d(u,v),dst(x,y)]$. Our Experimental results indicate that a linear support function is adequate. It should be noted that in Prazdny's algorithm for stereopsis [69], an exponential support function is used. The support function is a quantitative expression for the notion of local smoothness. Prazdny's choice of support function is intuitive, based on psychophysicaJ data. The same justification applies to motion correspondence. As an example consider the following exponential support function, which we used in our experiments:

$$S(d,l) \quad \frac{1}{l}\exp\left(-\left(\frac{d}{l}\right)^2\right)$$

where $l = dst(x,y)$ and $d = dst(u,v)$. As mentioned previously, there seems to be no great advantage in using an exponential support function in preference to a linear one. The advantage of clustering is that, once the clusters have been determined, the parameters of the support function are

obtained. Thus once the maximum velocity difference (i.e. the diameter of the cluster) is known, the largest velocity gradient that should be allowed can be calculated. This is the ratio of the cluster diameter (i.e. largest linear distance across the cluster) and the diameter of the support window in image space. Suppose this ratio is K and $f(d,l) = \frac{1}{k}(* - \frac{d}{\cdot})$, then the linear support function is

$$S(d,l) = \begin{cases} f(d,l) & \text{when } f(d,l) > 0 \\ k) & \text{otherwise} \end{cases}$$

The algorithm 2.1 outlines the steps involved in the computation of the matches.

Algorithm 2.1: ***Finding motion correspondence by support disparity without clustering.***

```
begin
Fl := {X |X is a point with coordinate (x,y) on the first frame };
F2 := {X jX is a point with coordinate (x,y) on the second frame };

(* Computation of total support for each disparity *)

for each element, p of Fl do
  for each element,q of F2 within a radius, R of p do
    Totalsupport(pq) := 0;
    for each element, r of Fl do
      for each element, s of F2 within a radius, R of r do
        Support := support provided by vector rs to vector pq
      end_for,
      Totalsupport(pq) := Totalsupport(pq) -f Support;
    end_for;
  end_Jor,
end_for9
```

```
(* Finding correspondences from the total supports *)

for each element p of Fl do
  for each element q of F2 within a radius,R of p do
    Find Maximum(Totalsupport(pq))
   (* the vector pq corresponding to Maximum(Totalsupport(pq)) gives
    the correspondence *)
  end^Jor,
end_for;

end (* algorithm 1 *)
```

2.4.2. Retinal motion detection with velocity clustering

The simple algorithm presented above works well in most instances. However, for cases where there are a large number of match possibilities for every point, the method is cumbersome. In such instances, a separate layer of space unspecific displacement (or motion) units are computed. This is like a cluster space of retinal motion parameters (i.e. *u*,v) with a spatial SRP that extends over the window of the image that is of interest. Each unit in this cluster space collects "votes" or support from the location specific displacement (or match) vectors of identical magnitude and orientation, from the layer below.

The clustering approach to visual motion measurement and segmentation has a number of attractive features that will be mentioned shortly. The clustering process is best understood in terms of partitioning of graphs. Let $G = (V, E, W)$ be a weighted undirected graph with vertex set V, edge set E and a distance or weight function $W: E \rightarrow R^2$ (the set of

nonnegative reals). A partition of the set of vertices into k sets $(C_1, C_2, .., C_k)$, is called a k-split. The sets C_i are called *clusters*. In addition there is an *objective function* $\psi: C_1, .. C_k \rightarrow R^+$ defined on the k-split.

The clustering problem can now be defined in two ways. In case it is known a priori how many clusters (i.e. k) are present then:

Definition I: Given a graph G, an objective function f and an integer k, find the k-split which minimizes the objective function. In other words, find $(C_1^*, ..., C_k^*)$ that

$$\psi(C_1^*, .., C_k^*) = \min \{\psi(C_1, .., C_k) \mid (C_1, .., C_k) \text{ is a } k\text{-split for } G\}$$

Under some circumstances however, the number of clusters are not known a priori. In this case one can specify a threshold θ, whence the clustering process is defined as:

Definition II: Given a graph G, an objective function ψ and a positive real valued threshold θ, find for the least value of k, a k-split with objective function value $\leq \theta$.

Clustering can also be defined with respect to an n-dimensional feature space in an exactly analogous manner. Of course in this case one must formulate a distance metric for points in the feature space. An alternative to distance or dissimilarity function is a similarity function, examples of which were cited in the previous section. In the case that similarities are used in place of distances, the objective function is usually maximized. Note that the definition does not require the distance/similarity to be metric.

The objective function V plays a crucial role in clustering. As mentioned before the motion vectors and their spatial positions form elements that belong to a four dimensional space. It is expensive to compute clusters in this four dimensional space. Conceptually, this is an attractive framework in which to view the motion measurement and segmentation problem. The algorithm in the previous section is similar to the stereo algorithm proposed in [69], On the other hand, the cluster based formulation *unifies the notion of matching and segmentation.* The assumption is that with adequate data from all the different surfaces moving in the visual field separate significantly large clusters (compared to random fortuitous clusters) will indicate the corresponding motion segment* Then, the desirable matches belong to one of these larger clusters while mismatches are scattered into small noise clusters. In order that the clusters be well rounded and not in the form of "stringy chains", the objective function must be chosen with care.

A good clustering strategy is provided by the so called complete linkage or furthest neighbor technique. The objective function in this case is defined to be the largest distance between pairs of elements computed over all pairs that belong to the same cluster. Unfortunately clustering with this function is known to be NP-hard for feature spaces of dimension two or more, when there is more that one cluster [75].

To overcome the problem of large dimensionality and computational complexity, the method adopted was to project the four dimensional feature

space into a two dimensional space of motion vectors without spatial indexing. Furthermore, simulation with various synthetic data showed that the problem of chain formation showed up vary rarely in the projected space. For this reason an agglomerative hierarchical clustering strategy was adopted. Initially all the elements belong to singleton clusters. At each stage of processing (for simplicity assume sequential execution of the stages), each cluster merges with its nearest neighbor. The process continues until there is only one cluster. This is called the single linkage method and essentially generates a *minimal spanning tree* of the feature space graph, in which the nodes are the elements to be clustered and the edges are the distances between them.

The computational complexity of the algorithm is $O(n^2 \log n)$ for the serial case with n elements. The algorithm can be easily implemented in parallel with p processors with a complexity of $O(\frac{n^2}{p} \log n)$ [76]. Another advantage of adopting the clustering view is that a number of suboptimal algorithms have been published and could be used for this application (e.g. see [37] for a factor 2 $O(nk)$ algorithm).

The implemented program follows an algorithm given in [25]. The clustering metric is the Euclidean distance between two motion vectors. The number of clusters depends upon a threshold for the similarity. This threshold is chosen depending on it stability, meaning that small changes to

it should not affect the clustering in any significant way. The cluster trees (or dendograms) can be seen for the synthetically generated data of two differently moving surfaces in figure 2.8.

The clusters so formed now compete against each other and only the larger clusters, i.e. the ones with accumulated votes in the same order of magnitude, are kept These clusters then mediate the matching process in the lower level of displacement vectors (figure 2*4). By this process two things are achieved:

1. Noise points and spurious matches are avoided*
2. In the case of multiple body motion, the clusters provide a convenient label for segmenting the displacement field.

An outline of the algorithm follows:

Algorithm 2.2: *Finding motion correspondence by clustering followed by application of support disparity.*

```
begin

FR1 := (X |X is a point with coordinate (x,y) on the first frame }

FR2 := {X |X is a point with coordinate (x,y) on the second frame }


(* setting up the table for clustering *)

for each element,p of Fl do
  for each element,q of F2 within a radius,R of p do
    displacementjjc_direction := (x coordinate of q) - (x coordinate of p);
```

```
    displacement_y_direction := (y coordinate of q) - (y coordinate of p);
    clustertable[displacement_x_direction , displacement_y_direction] :=
     clustertable[displacement_x_direction , displacement_y_direction] + 1;
  end_for;
end_for;

Find the clusters in the two-dimensional array clustertable;
Remove clusters with weak overall support (votes);
From the clusters find the feasible disparities;
Consider points in the feasible disparity ranges only, and apply Algorithm 1;

end.(*algorithm 2*)
```

The matching algorithm is formulated according to whether the points are labeled or not. In case of unlabeled points (as in the above algorithms) :

> All neighboring points support (vote for) a particular disparity value. Similar values support each other strongly in a local region. Shorter length disparities are preferred. A point adopts a match for which it finds the maximum support.

The strategy is similar in spirit to the more sophisticated matchers, for instance, those using labeled points (e.g. [66]). The feature points can carry labels which are computed from the outputs of the nine basis operators. A label is a code that identifies the image point in question. Now the matcher weights the "supporting" votes according to the similarity of these codes. However, we avoid iterative refinement, which is usually employed in similar algorithms [14].

2.5. Experiments

Synthetic Images: Experiments have been conducted on synthetic images of spheres and planes "painted" with random dot patterns. All the objects are opaque and sometimes intersect each other. The choice of spheres and planes is motivated by the necessity of local smoothness in the imaged depth. Yet, at the same time, since there are multiple differently moving objects as well as occlusion of one body by another, motion boundaries do occur.

The image formation technique was the perspective projection. In a single image there could be one or more instances of the above primitive objects moving with similar or different velocities (translational and rotational) in 3-space. An illustrative depth map of the surface of a sphere embedded in a plane is shown in Plate 2.I.

For single body motion the matches were found with close to 100% accuracy. Addition of uniformly distributed uncorrelated noise points to a level of 10% did not cause any significant difference in the level of correct matches found. However, the noise points generated some spurious matches among themselves. The clustering approach works better in this situation with considerable removal of noise points and false matches.

As a conservative estimate the average number of plausible match vectors considered was of the order of 10 - 15. Of course in regions with dense random dot patterns this number was more. Even with larger

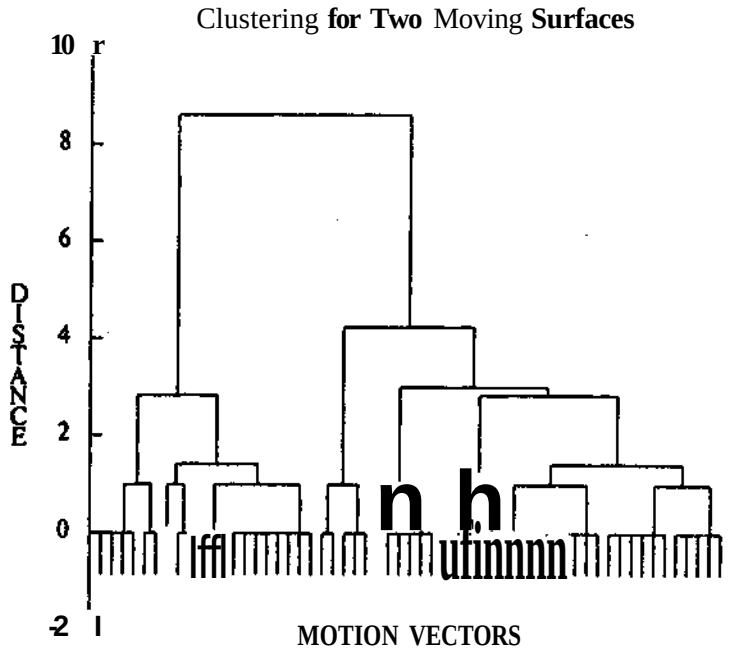


Figure 2.8 Cluster Tree for two body Motion

numbers the selection of the correct match was possible with the support disparity approach. These figures for plausible match vectors fell to a third of their number with the clustering approach. There was also a speedup of execution by a factor of around ten.

With two body motion about 94% of the correct matches could be obtained. The hardest matches to find lay on the border of the two bodies, as was to be expected. The dendogram (cluster tree) is shown in figure 2.8. Also, from Plate 2.IV it can be observed that the matches have been found correctly almost everywhere (comparing with Plate 2.II). The exception occurs at the boundary of the two bodies, where incorrect matches were found. An intensity coded view of the cluster generated for this case is

shown in Plate 2.III. The pixel positions and intensity represent the location (in velocity space) and population of the relevant cluster (or bin). In the case of totally transparent bodies the algorithm's performance is drastically reduced with 50 -60 % wrong matches being found.

Images of Natural Scenes: Quantitative justification of performance is difficult on natural scenes. Through manual inspection it has been found that the number of wrong correspondences obtained are insignificant. Plate 2.V demonstrates the result of applying the algorithm on a natural scene. The top right box depicts a single snapshot (frame) of the scene, while the bottom right box shows the interest points computed, superposed on the scene. The box at bottom left illustrates the computed image motion vectors, which can be compared with the manually computed vectors. The latter were obtained by selecting the points to match from two frames by inspection and then matching them as they were selected manually. In Plate 2.VI, points from two successive time frames are shown superposed, to illustrate the input available to the clustering algorithm. Here the coding of the points from different frames is done by bright and dim dots. Finally the cluster obtained for this image sequence is shown in Plate 2.VII. However, correct matches associated with roughly 40% of the points have been found. This discrepancy is a result of the uncertainty associated with the determination of interest points.

It is estimated that even with some amount of input/output processing, the part of the algorithm which could be parallelized took about 75% of the time on a serial processor. With the removal of file manipulations this could rise to over 95% or more of the time. It is feasible to implement the algorithm on a 128-processor MIMD machine (BBN Butterfly) with considerable improvement in running time.

2.6. Conclusions

The goal of the research here, was to formulate a computational framework for the measurement of retinal motion. It was desirable that the motion measurement algorithm be implementable in parallel, and conform to a connectionist implementation strategy. An important consideration was graceful degradation in the presence of increasing amounts of noise, and the ability to handle multiple moving objects. An important issue, relating to the task of retinal motion measure is the choice of the matching primitive or token and the process for obtaining these primitives in an image. The overall framework of the algorithm is based on the matching paradigm of motion perception. This is based upon the belief that some form of matching, either involving spatio-temporal gradients or other feature primitives, is essential to solving the motion perception problem. This paradigm is by no means inviolable, as has been shown recently in [3], for certain imaging situations.

The work with synthetic images served to lay the preliminary groundwork for evaluating the proposed matching algorithm. The success of this study showed that the scheme is reliable enough to test on natural images. Of course the heart of the matter is to be able to determine the interest points without elaborate processing. Thus experiments with natural images was thought to be contingent upon being able to formulate and compute feature primitives that are stable and recoverable with local operators. The orthogonal decomposition operator described here (see also [12]) proved adequate for the purposes of applying the clustering algorithm on natural images. This "interest" operator is simpler than other corner finding algorithms like the ones described in [27, 51, 62], although its performance is comparable to the best of them ([80])* Incidentally, the operator described here was also used recently in another image processing context with considerable success (see [4]).

To summarize, a list of the salient advantages of our algorithm is given below:

1. Applicable to multiple moving objects.

2* Good behavior in the presence of noise.

3. Automatic segmentation **for** image areas projected from different moving objects or parts of the same rigid surface differentiated by sharp depth changes.

4. The clustering formulation proposed, is a mathematically well defined paradigm for motion segmentation. Furthermore, the complexity of this approach is well understood and efficient suboptimal algorithms can be used.

5. Conceptual simplicity and amenability to parallel implementation.

6. Matching and segmentation are handled uniformly, under the same paradigm.

PLATE I. Example of a synthetically generated surface

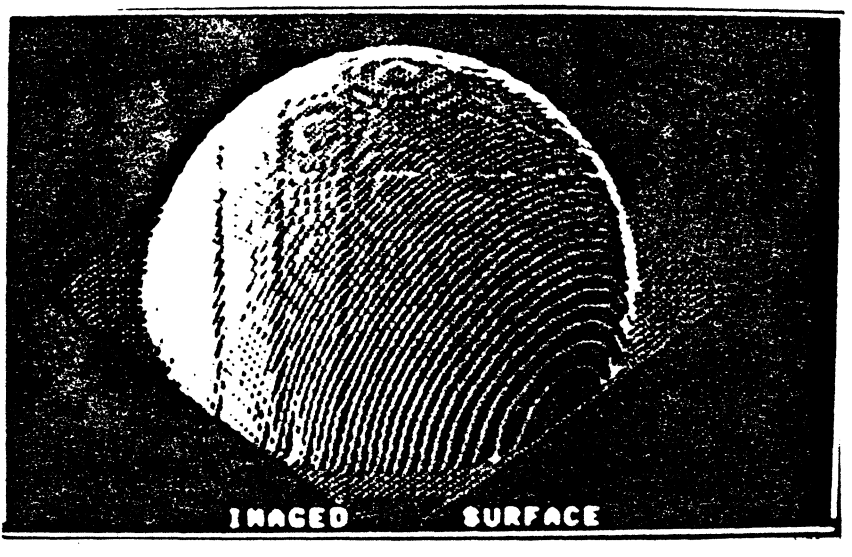


PLATE 2.II. Correct matches for two body motion.

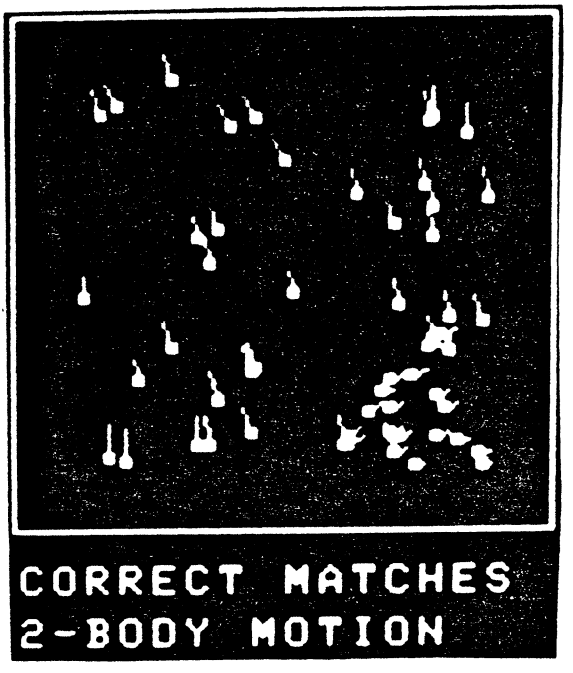


PLATE 2.III. Cluster for two body motion.

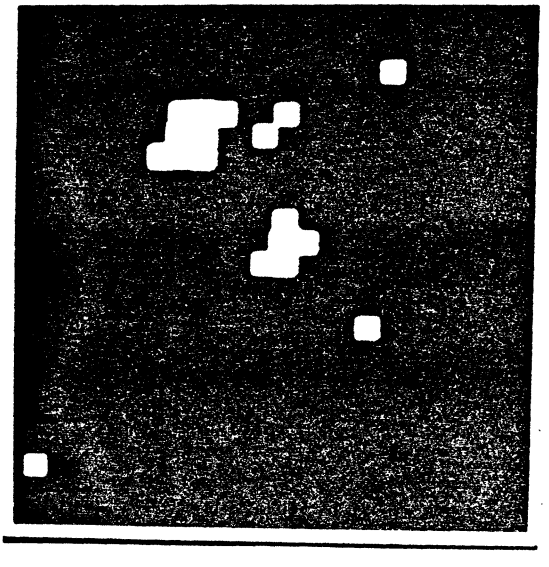


PLATE 2.IV. Computed two body motion matches.

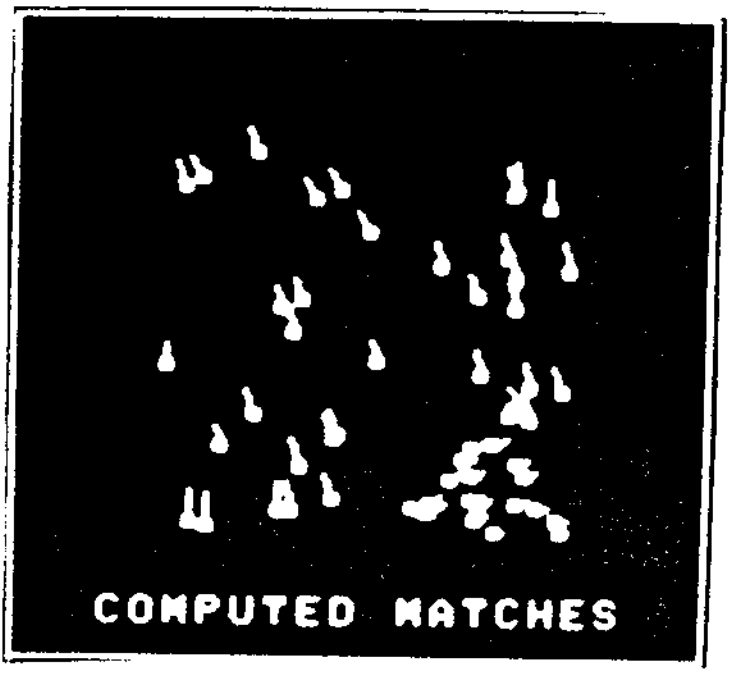


PLATE 2.V. Results obtained on a natural image
(i) computed vectors bottom left
(ii) manually determined vectors top left
(iii) interest points from 1st frame at bottom right

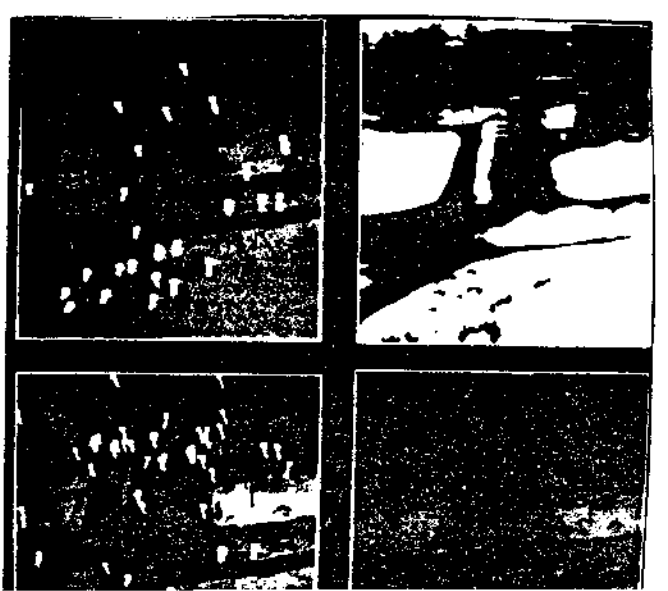


PLATE 2.VI. Interest Points from two consecutive frames shown superposed (light and dark points).

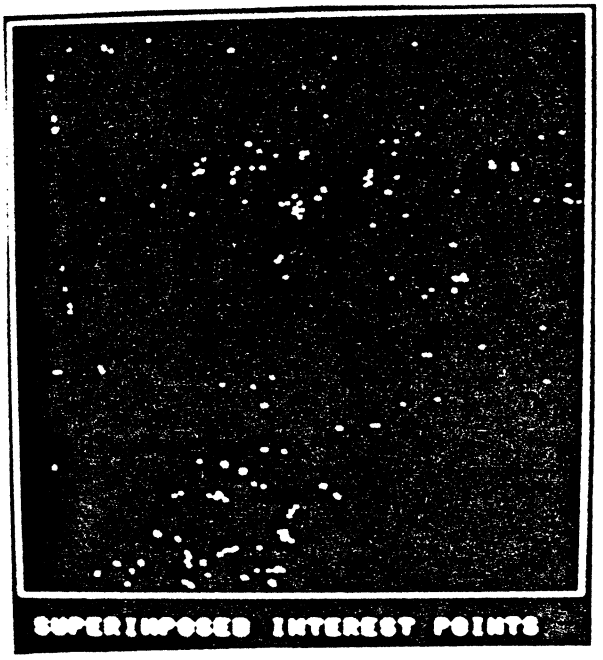
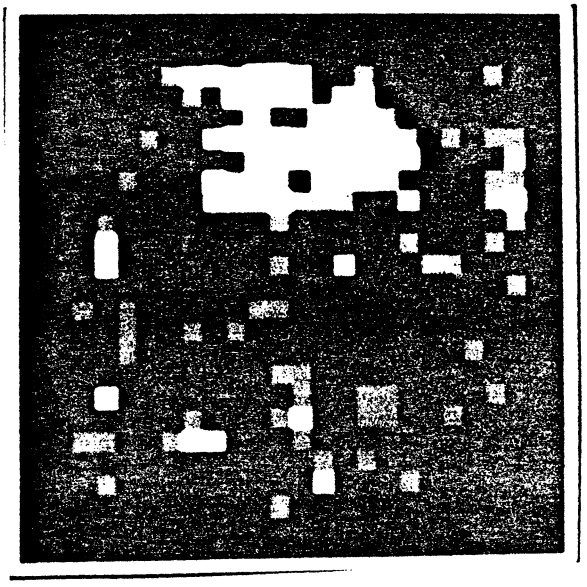


PLATE 2.VII. Cluster for the natural image.



Chapter Three

Physical Constraints on Image Motion

3.1. Introduction

This chapter establishes a mathematical framework for investigating the motion perception problem, with a view to understanding the adequacy of the resultant mathematical constraints. The reader who is knowledgeable in the basics of the area, can start with the last section of the chapter, which summarizes the contribution towards theoretical understanding of the mechanics of motion interpretation.

The motion of a body can be characterized by the rate of change of the positions of various points on its visible surface. Instantaneously, this corresponds to a three dimensional velocity field. If the body (or surface) is rigid, then, this velocity field can be described by the set of three dimensional position coordinates and *six global parameters* (see figure 3,1), which are:

(i) The three components of the velocity of any point O on the body. These are called the translation parameters.

(ii) The rotational velocity components of a coordinate frame, with origin O, attached rigidly to the body.

A standard result from kinematics and geometry (see [21]) is that although the rotational parameters are invariant with respect to the choice of the origin O, of the body frame, the translation parameters are dependent on the choice of O.

In general, computing three dimensional motion from monocular two dimensional image motion flux is an underdetermined problem, admitting an infinite number of solutions. However, most of the moving objects in our environment are rigid, and the *rigidity constraint* greatly simplifies the task of representation and analysis of visual motion [86]. From a practical standpoint, the study of rigid body motion is interesting, since it finds widespread applications in the areas of optical navigation, tracking and recovery of 3D structure of rigid objects. The following analysis explores the ramifications of this central assumption in the interpretation of three dimensional structure and motion from two dimensional image motion (see Ullman's paper [88] for a discussion of nonrigid motion perception).

The previous chapter introduced the notion of *optical flow* as an abstraction for image motion. It is a fact that, as yet, it seems very difficult to compute optical flow. However, it can be estimated by spatio-temporal

interpolation from discrete displacement measurements, when the sampling is "adequate". Under this adequate sampling assumption, the following analysis deals with the optical flow representation for image motion.

When considering motion of rigid bodies, there are two cases of interest, namely, *egomotion* and *general motion*. Egomotion or self-motion refers to the movement of the camera or sensor in a static environment. The image flux, or *optical flow*, generated due to such a motion is due to a single relative movement, i.e. between observer and static environment. In contrast, general motion implies that there is more than one object moving with different velocities in the observers field of view. In this case the optical flow field consists of many *segments* corresponding to the various moving surfaces. Each segment is characterized by the translational and rotational velocities of the associated moving rigid surface inducing the optical flow. These velocities are called the parameters of motion for the rigid surface.

The rigid motion parameters are usually expressed with respect to a frame of reference attached to the moving surface, which is assumed to coincide with the observers frame of reference at the time of observation. The problem is to determine the motion parameters corresponding to a optical flow field segment. If the depth of the scene is unknown then it can be shown that only the rotation - which is depth invariant - can be determined uniquely; whereas the three translation parameters can only be

determined up to a scale factor (this is the depth scaling effect). Thus we can determine five parameters to characterize the motion in this case.

The concern here is with the physical constraints that make it possible to compute the five parameters of rigid motion and the structure of the moving surface from retinal stimulus such as optical flow.

The optical flow field comprises two parts, corresponding to the rotation and the translation, respectively, of the inducing motion, and is constrained at every point by the parameters of the motion. Motion perception becomes simpler in the instances when the optical flow field can be computationally separated into the respective components [39]. A familiar illustration of this is the case of motion parallax observable at depth discontinuities in the retinal field. The effect is to reduce the dimensionality of the space of unknowns. Unfortunately, this seems to be very hard to accomplish, in general. Motion parallax is the basis for an algorithm by Lawton [70]. Other approaches to the problem can be found in [19, 52], involving nonlinear least square techniques or using local constraints involving derivatives of the optical flow.

Algorithms for rigid motion perception are difficult to design due to two main reasons:

(1) The space of parameters is of high dimensionality (e.g. five).

(2) The constraint equations obtained by optical flow measurements are non-linear.

There have been some clever implementations of non-linear search algorithms to interpret 3D motion from optical flow data [67, 68]. There have also been discrete point tracking algorithms by Tsai and Huang [84] and Fang and Huang [28, 29] and Longuet-Higgins [53]. In some of the latter algorithms, the nonlinear motion equations are linearized in terms of synthetic parameters, which are nonlinear combinations of the actual motion parameters. Tsai and Huang, and Fang and Huang, note the cases when such algorithms fail to compute motion parameters.

An important aspect of the following analysis will be the examination of situations where the monocular optical flow field could be interpreted in more than one way. An instance of such ambiguity is the optical flow field due to motion of a plane [82].

A geometric analysis of the problem of computing 3D motion parameters from 2D image velocities has been done by Longuet-Higgins and Prazdny [52]. The constraint equations that they derive are simple in form, but deal with velocities. To implement a motion analysis algorithm based on these equations, one makes the assumption that the temporal grain of the observations is fine enough to talk meaningfully about the *velocities* or time derivatives of both the image and world positions. Representing motion by velocity parameters entails making a first order approximation of

the temporal behavior associated with the motion. Thus, for example, if the displacement of a particle moving in one dimensional space is Δx in time Δt, then $\frac{\Delta x}{\Delta t}$ is a good approximation for the velocity only when Δt is small enough such that the change in velocity in this time period is small.

An alternative derivation is due to Tsai and Huang [84]. Their approach is to analyze the relation between the projected displacement vectors in the image plane due to an arbitrary rigid displacement of a set of points in 3D. It is known [21] that this type of motion can be characterized by a rotation about an axis passing through the origin of the reference coordinate frame and a translation.

The assumptions underlying the work reported here are:

(i) The motion being observed, is due to a rigid surface.

(ii) The time constant (or sampling interval) of the sensor is small enough to make a first order approximation of the temporal behavior due to the motion being observed.

3.2. Review of related work in the analysis of motion geometry

The computation of rigid motion parameters from image displacement vector fields has been studied by a number of researchers. Egomotion has been considered by Longuet-Higgins and Prazdny [52], Prazdny [67], Waxman and Ullman [89] and Bruss and Horn [19]. Longuet-Higgins and Prazdny examine ways of determining 3D structure and motion parameters

from optical flow, given an accurate reconstruction of the optical flow field. They show that for non planar surfaces *local analysis* of the flow field yields a cubic constraint involving the motion parameters. Prazdny ([67]) has devised a five point algorithm to solve for the motion parameters from nonlinear constraint equations. Waxman and Ullman's method depends upon reconstruction of the optical flow field analytically, in local neighborhoods. Bruss and Horn propose a least square solution to the parameter estimation problem.

Some other computational approaches attempt to segment the optical flow field into translational and rotational components, albeit approximately. An example is the method of Rieger and Lawton [70] where the change of rotational flow at steep depth gradients, is treated as noise. Jain [48, 49] computes the *focus of expansion* before computing the image displacements and uses the former to guide the correspondence for finding the latter.

All the above methods compute the motion parameters from optical flow, i.e. continuous or differential image motion. An alternative approach is to consider evaluating the motion parameters and 3D structure from discrete point correspondence. Ullman [86] shows that three views of four non coplanar points is adequate to determine the structure and motion of these points under orthography. Tsai and Huang [84] prove that the motion of seven points not lying on two planes, one of which passes through the origin, nor on a cone passing through the origin, can be uniquely computed,

from discrete displacements. Fang and Huang [28, 29] prove that structure and motion of nine points not lying on a second order surface passing through the origin is uniquely determined from image displacements. Nagel and Neuman [60] and Roach and Aggarwal [71] have also looked at the problem of determining motion from discrete displacements.

Yet another approach to the problem of motion parameter computation has been to restrict the motion to simplify the analysis. Webb and Aggarwal [92] Hoffman and Flinchbaugh [41] and Hoffman and Bennett [43] analyze rigid motion with the additional assumption of fixed axis of rotation or planarity. An major motivation for this type of analysis is that, it models the locomotion of man and animals.

3.3. The Geometry of Rigid Motion

Consider a sensor moving relative to a static scene. The coordinate frame (X,Y,Z) is fixed to the sensor (see figure 3.1). The viewing direction is along the positive z-axis.

Under orthography, the projection equation relating the position of a point in three space $P=(X,Y,Z)$ to its image $p=(x,y)$ is:

$$(x,y)=(X,Y)$$

Under perspective projection, the "image" is formed by "rays" from points in three space (i.e. world points). These rays are constrained to pass through a nodal point called the center of perspectivity. The imaging

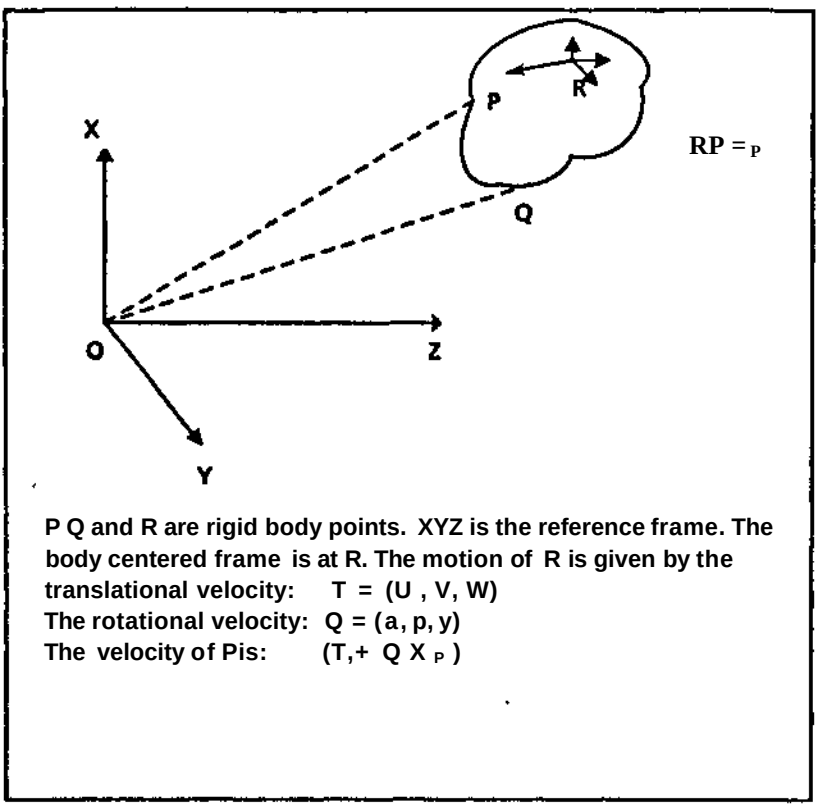


Figure 3.1 Representation of rigid motion parameters

geometry is shown in figure 3.2. The nodal point is O, which is also taken as the origin of the frame of reference. An image point p = (x,y) corresponds to the world point P = (X,Y,Z). Here the *focal length* of the imaging system is f.

The equation of the ray **OP** is :

$$\frac{X}{x} = \frac{Y}{y} = \frac{Z}{f}$$

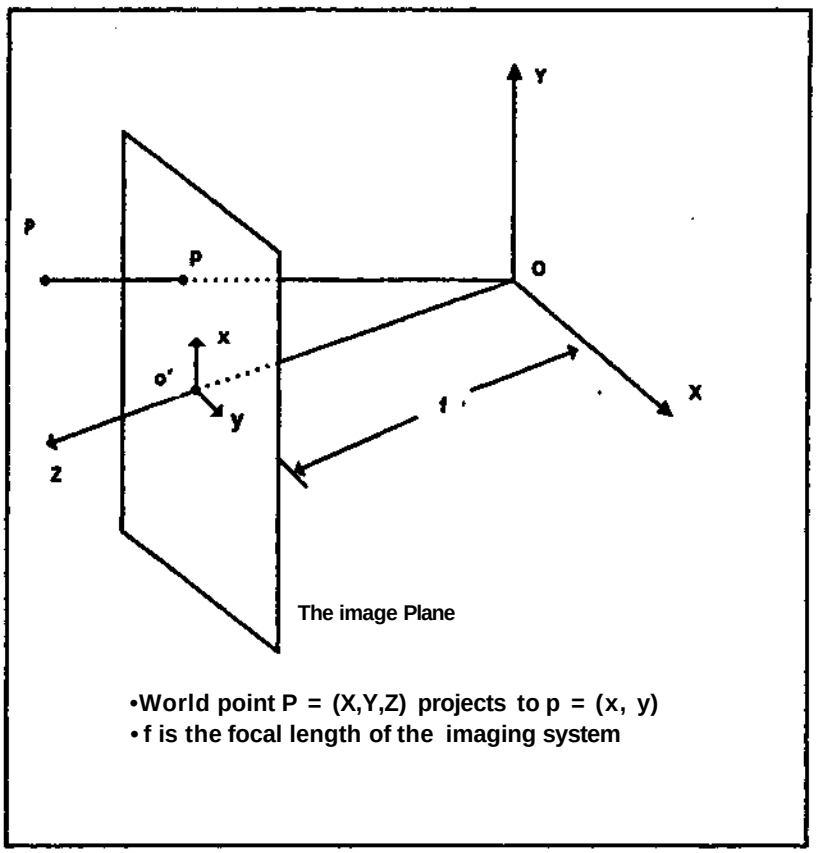


Figure 3.2 Perspective Imaging Geometry

$$x = \frac{fX}{Z}; \quad Y = \frac{fY}{Z}$$

The above projection is denoted by *(X,Y,Z)-+(x,yJ).* Similarly, the projective relation between another world point P^1 and its image is $(X', Y', Z') \rightarrow (x', y', f)$

A rigid body is defined as a set of points whose relative euclidian distances from all other points in the set are invariants with respect to the transformations of rotation and translation. In addition, since we will generally deal with opaque objects and hence will observe points on a

surface (or a manifold) in 3 space. In other words the 3 cartesian coordinates of of a point on a rigid body are not independent. Formally,

where

$$\pi = \{(X,r,Z) \mid \textit{point on the surface of } B\}$$

$$f(X,Y,Z) = 0$$

When the body B is displaced with respect to the frame of reference, we obtain a new representation

The displacement is described by the affine transformation

$$X^* = [i?]X + T \qquad (3\text{-}1)$$

Any displacement of a rigid body can be modeled by the above equation, which describes a rotation about an *axis through the origin* and a translation specified by the vector T.

The rotation matrix is orthonormal and its determinant is unity. Since any matrix can be expressed as the sum of a symmetric and a skew symmetric matrices uniquely we have:

The axis of rotation is denoted by the unit vector $(l,m,n)_y$ whose components are the direction cosines of the axis of rotation and $I^2 + m^2 + n^2 = 1$. The angle of rotation about this axis is $. Then, we have:

$$R_{sym} = \begin{bmatrix} l^2 + (1 - l^2)\cos\theta & lm(1 - \cos\theta) & nl(1 - \cos\theta) \\ lm(1 - \cos\theta) & m^2 + (1 - m^2)\cos\theta & mn(1 - \cos\theta) \\ nl(1 - \cos\theta) & mn(1 - \cos\theta) & n^2 + (1 - n^2)\cos\theta \end{bmatrix}$$

$$R_{skew} = \begin{bmatrix} 0 & -n & m \\ n & 0 & -l \\ -m & l & 0 \end{bmatrix} \sin\theta$$

If the rotation angle is small with respect to the precision of retinal measurement, the rotation matrix can be written in terms of the three component rotations about the individual axes [45]. In this case R and T are given by

$$R = \begin{bmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{bmatrix} \qquad \mathbf{T} = \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}$$

where $\omega_x = l\theta$, $\omega_y = m\theta$ and $\omega_z = n\theta$. Substituting for R and T in equation (3.1) we have,

$$X' = X - \omega_z Y + \omega_y Z + t_x \qquad (3\text{-}2.1)$$

$$Y' = Y + \omega_z X - \omega_x Z + t_y \qquad (3\text{-}2.2)$$

$$Z' = Z - \omega_y X + \omega_x Y + t_z \qquad (3\text{-}2.3)$$

or,

$$\Delta X = t_x - \omega_z Y + \omega_y Z \qquad (3\text{-}3.1)$$

$$\Delta Y = t_y + \omega_z X - \omega_x Z \qquad (3\text{-}3.2)$$

$$\Delta Z = t_z - \omega_y X + \omega_x Y \qquad (3\text{-}3.3)$$

where,

$$\Delta X = X' - X \qquad \Delta Y = Y' - Y \qquad \Delta Z = Z' - Z$$

We define the parameter vector **a** for characterizing the motion, where

$$\mathbf{a} = (t_x, t_y, t_z, \omega_x, \omega_y, \omega_z)^t$$

Motion perception involves the recovery of the parameters of motion, as well as the structure (or shape) of the moving object. The geometric properties of the three dimensional surfaces and points are related to the geometry of their image. Thus the projective transformation involved in the image formation process must be analyzed. The subsequent analysis considers both the cases of perspective as well as orthographic projections.

3.4. Motion under Orthography

When the model of image formation involves orthographic or parallel projection, then the mathematical formulation of the problem becomes considerably simpler. It can be argued that this is a valid model of image formation when viewing distant objects, or when the focal length of the camera is large compared to the distance of the viewed surfaces, or when the viewing area is small and centered around the line of sight - as in the case of the field of view corresponding to the *fovea* in the retina. Consider an image point $p = (x, y)$ projected by the world point $P = (X, Y, Z)$. Assuming that after a short while the point moves to a position given by $P' = (X', Y', Z')$ while its image moves to $p' = (x', y')$ the following relations are obtained from equations (3.3):

$$\begin{aligned} \Delta x &= x' - x = \Delta X = t_x - \omega_z Y + \omega_y Z \\ \Delta y &= y' - y = \Delta Y = t_y + \omega_z X - \omega_x Z \end{aligned} \tag{3.4}$$

Optical flow is the time derivative of the image position vector and is denoted by (u,v) where

$$(u,v) = (\dot{x},\dot{y}) = (\dot{X},\dot{Y})$$

Alternatively,

$$u = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt} \qquad v = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \frac{dy}{dt}$$

The motion parameters are now the translational velocity $\mathbf{V_T} = (U,V,W)$ and the rotational velocity $\Omega = (\alpha,\beta,\gamma)$ where:

$$U = \lim_{\Delta t \to 0} \frac{t_x}{\Delta t} \qquad V = \lim_{\Delta t \to 0} \frac{t_y}{\Delta t} \qquad W = \lim_{\Delta t \to 0} \frac{t_z}{\Delta t}$$

and

$$\alpha = \lim_{\Delta t \to 0} \frac{\omega_x}{\Delta t} \qquad \beta = \lim_{\Delta t \to 0} \frac{\omega_y}{\Delta t} \qquad \gamma = \lim_{\Delta t \to 0} \frac{\omega_z}{\Delta t}$$

therefore the equations relating image and 3D motion are

$$\begin{aligned} u &= U + \beta Z - \gamma y \\ v &= V - \alpha Z + \gamma x \end{aligned} \tag{3.5}$$

These equations are exactly identical in form to those obtained under the discrete case (assuming small rotation), i,e. equation (3.3). Strictly speaking, according to the nomenclature adopted before, the motion parameters for the discrete case are $(t_x,t_y,\omega_x,\omega_y,\omega_z)$ and those for the differential case are $(U,V,\alpha,\beta,\gamma)$. However, since equations (3.3) and (3.5) are identical in form, all subsequent analysis is based on the latter equation. Furthermore, the parameters (it will be evident later that only the rotational parameters are of interest here), in both the differential as well as

the discrete cases will be referred to by the symbols (α,β,γ). The treatment of both the cases is identical, the only difference being that derivatives in the differential analysis correspond to differences in the discrete case.

3.4.1. On the information available in the optical flow field

Observe from equation (3.5) that the image displacement (or image motion field) consists of a translational part and a rotational part. The translational motion parameters are dependent on the origin of reference. In fact the parameters, intrinsic to the motion are those of rotation. Thus relative to a particular point, say the origin (0,0), equation (3.5) becomes:

$$\begin{aligned} u &= \beta Z - \gamma y \\ v &= -\alpha Z + \gamma x \end{aligned} \tag{3.6}$$

where u actually means $u - u(0,0)$, v is $v - v(0,0)$ and Z is $Z - Z(0,0)$. It should be emphasized here that Z denotes depth *relative* to a certain point of reference (in this case it is the origin). If the structure or relative depth is not known then the parameters (α,β,γ) are not completely recoverable. There is an exact analog of equation (3.6) for the discrete case, obtainable from equation (3.3).

Proposition I When the depth function (or structure) is *non planar* the following parameters are uniquely determined from the image displacement field:

(1) The rotation about the axis aligned with the line of sight, i.e. 7.

(2) The ratio of the other two parameters, i.e. -^-.

p

Proof: The proof is by contradiction. Consider the motion of the *non planar* surface Z_u which is described by the parameters *(ctufluiii).* The image motion equations (from equation (3.6)) are:

$$\begin{aligned} u &= \beta_1 Z_1 - \gamma_1 y \\ v &= -\alpha_1 Z_1 + \gamma_1 x \end{aligned} \qquad (3.7)$$

Now suppose there is another surface Z_2> whose motion is characterized by the parameters (<*2,£2,72), such that the image motion field in both the cases is the same. The motion equation for the second surface is:

$$\begin{aligned} u &= \beta_2 Z_2 - \gamma_2 y \\ v &= -\alpha_2 Z_2 + \gamma_2 x \end{aligned} \qquad (3.8)$$

Furthermore, the following relations hold:

$$\begin{aligned} A\gamma &= \gamma_1 - \gamma_2 \\ \frac{\alpha_1}{\beta_1} &= \frac{\alpha_2}{\beta_2} \end{aligned} \qquad (3.9)$$

From equations (3.7) and (3.8); the following relations are obtained: (3.10)

$$-\alpha_2 Z_2 + \alpha_1 Z_1 + A\gamma x = 0$$

Now since $\alpha_2/\beta_2 = \alpha_1/\beta_1$: $$Z_1 = \frac{-\alpha_2 \Delta\gamma\, y + \beta_2 \Delta\gamma\, x}{\alpha_2\beta_1 - \alpha_1\beta_2}$$

But this is contrary to the assumption that Z_x is non planar. Therefore:

$$\frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2}$$

Again, this implies (considering equation (3.10)) that

$$\Delta\gamma = 0 \qquad or \qquad \gamma_1 = \gamma_2$$

This completes the proof of Proposition I.

Proposition II. The image displacement field generated by a planar surface is linear in the arguments (x,y). In addition, the parameters $\frac{\alpha}{\beta}$ and γ are uniquely determined by the image displacement field if and only if $\alpha p + \beta q = 0$, where (p,q) is the gradient of the planar surface.

Proof: Consider the equation of the planar surface $Z(x,y)$:

$$Z = \bar{p}x + \bar{q}y + \bar{d}$$

If the motion of the surface is characterized by the parameters $(\bar{\alpha},\bar{\beta},\bar{\gamma})$. The image motion (or optical flow) is given by:

$$\begin{aligned} u &= \bar{\beta}(\bar{p}x + \bar{q}y) - \bar{\gamma}y \\ v &= -\bar{\alpha}(\bar{p}x + \bar{q}y) + \bar{\gamma}x \end{aligned} \tag{3.11}$$

The above equation indicates that for planar surfaces the optical flow is linear. It is also true that when the optical flow is linear then the moving surface is planar. Now considering equation (3.6) and substituting for (u,v) from equation (3.11) and rearranging terms:

$$\begin{aligned} 0 &= (\bar{\beta}\bar{p} - \beta p)x + (\bar{\beta}\bar{q} - \beta q - \bar{\gamma} + \gamma)y \\ 0 &= (-\bar{\alpha}\bar{p} + \alpha p + \bar{\gamma} - \gamma)x + (-\bar{\alpha}\bar{q} + \alpha q)y \end{aligned} \tag{3.12}$$

Since the above equations are valid for the entire image we have:

$$\beta p = \bar{\beta}\bar{p} \tag{3.13.1}$$

$$\beta q - \gamma = \bar{\beta}\bar{q} - \bar{\gamma} \tag{3.13.2}$$

$$ap - 7 = \bar{a}\bar{p} - \bar{7} \tag{3.13,3}$$

$$aq = 5y \tag{3.13.4}$$

Eliminating p,q and 7 from the above:

$$\beta\left(\frac{\bar{\alpha}\bar{q}}{\alpha}\right) - \alpha\left(\frac{\bar{\beta}\bar{p}}{\beta}\right) = (\bar{\beta}\bar{q} - \bar{\alpha}\bar{p})$$

or

$$\cdots$$

where $/* = \frac{\alpha}{\beta}$. The above quadratic equation has a unique solution if and only if:

$$(\bar{f}\bar{i}\bar{q} - \bar{o}\bar{p}'')^2 - A\bar{o}\bar{f}\bar{p}\bar{p}q = \{J\bar{q} + \bar{a}\bar{p}'')^2 = 0$$

Under this condition:

$$\frac{a}{\beta} = \frac{\bar{a}}{\bar{\beta}}, \qquad \frac{p}{q} = \frac{\bar{p}^*}{\bar{q}}$$

Therefore the image motion of planar surfaces uniquely determines the parameters $(\frac{-}{p}, 7, \frac{-}{9})$ if and only if $ap + 0q = 0$.

3.4.2. Summary **for the case of orthographic projection**

What we have shown is that:

(1) The analysis under orthographic projection for both differential and discrete motion are nearly identical.

(2) When the structure of the moving object is known, the motion parameters can be computed uniquely from image motion.

(3) When the structure is not known then the recoverable parameters are ($\frac{a}{a}$-,7,$\frac{p}{9}$). However in this case, the values are unique only when the moving surface is non planar, or a certain condition (see proposition II) holds*

3.5. Analysis of Rigid Motion for the Perspective Projection Model

Recall that the projective relation between image and world coordinate for a point P is given by,

$$x = \frac{fX}{Z}; \quad Y = \frac{fY}{Z} \tag{3.14}$$

The above projection is denoted by *(X,Y,Z)-*{x,y,f).* Similarly, the projective relation between another world point *P'* and its image is *{X',Y',Z')-*(x',y',f)* Following the schema used for orthographic projection we proceed to derive the equations of motion from first principles. Thus from equation (3.3) we have,

$$\Delta x = x' - x = f(\frac{X'}{Z'} - \frac{X}{Z})$$

$$\Delta y = y' - y = f(\frac{Y'}{Z'} - \frac{Y}{Z})$$

or,

$$\Delta x = f \frac{ZAX - XAZ}{Z(Z + AZ)} \tag{3-15.1}$$

$$\Delta y = f \frac{ZAY - YAZ}{Z(Z + \Delta Z)} \tag{3-15.2}$$

Recall that when the 3D rotation angles, characterizing the rigid motion, are "small" then the 3D displacement components are given by the relations:

$$\Delta X = t_x - \omega_z Y + \omega_y Z$$

$$\Delta Y = t_y + \omega_z X - \omega_x Z$$

$$\Delta Z = t_z - \omega_y X + \omega_x Y$$

Thus, substituting for ΔX, ΔY and ΔZ in the equation (3.15) we obtain an expression for the component of the retinal displacement,

$$\Delta x = f\frac{Z(t_x + \omega_y Z - \omega_z Y) - X(t_z + \omega_x Y - \omega_y X)}{Z^2 + Z(t_z + \omega_x Y - \omega_y X)}$$

or,

$$\Delta x = \frac{(ft_x - xt_z)/Z + f\omega_y - \omega_z y - \omega_z \frac{xy}{f} + \omega_y \frac{x^2}{f}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}}$$

similarly,

$$\Delta y = \frac{(ft_y - yt_z)/Z - f\omega_x + \omega_z x - \omega_x \frac{y^2}{f} + \omega_y \frac{xy}{f}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}}$$

The above equations express the the retinal displacement vector $(\Delta x, \Delta y)$ at an image point P = (x,y) in terms of the parameter vector **a** and the depth coordinate Z for corresponding world point p = (X,Y,Z). Another form of the above equations is,

$$\Delta x = \frac{(x_0 - x)\frac{t_z}{Z}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}} + \frac{f\omega_y - \omega_z y - \omega_z \frac{xy}{f} + \omega_y \frac{x^2}{f}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}} \qquad (3\text{-}16.1)$$

$$\Delta y = \frac{(y_0 - y)\frac{t_z}{Z}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}} + \frac{-f\omega_x + \omega_z x - \omega_x \frac{y^2}{f} + \omega_y \frac{xy}{f}}{1 + \frac{t_z}{Z} + \omega_x \frac{y}{f} - \omega_y \frac{x}{f}} \qquad (3\text{-}16.2)$$

where,

$$(x_0, y_0) = \left(\frac{ft_x}{Z}, \frac{ft_y}{Z}\right)$$

Note that, when the displacement is purely translational

$$\Delta x = \frac{t_x}{Z}(x_0 - x) \qquad (3\text{-}17)$$

This means that when the rotational component of the displacement is zero, the retinal displacement field converges to or diverges from a single point (x_0, y_0), ia the image plane. This point is called the *focus of contraction* (FOC) or the *focus of expansion* (FOE), depending on whether the translational motion is directed away from or towards the image plane (figure 3.3)•

From the retinal displacement field due to a particular motion, it is possible to estimate the parameters characterizing the motion. In addition, if the temporal sampling rate of our imaging process is high - meaning that the components of the displacement for a single time interval is small and the variable terms in the denominator of equation (3.16) are small compared to unity, i.e.

$$\frac{t_z}{Z} \ll 1 \qquad (A)$$

Given these assumptions, it is possible to derive the equations relating image motion to the motion parameters in the differential case. This is obtained by dividing equation (3.16) by a small time interval, A*, and taking

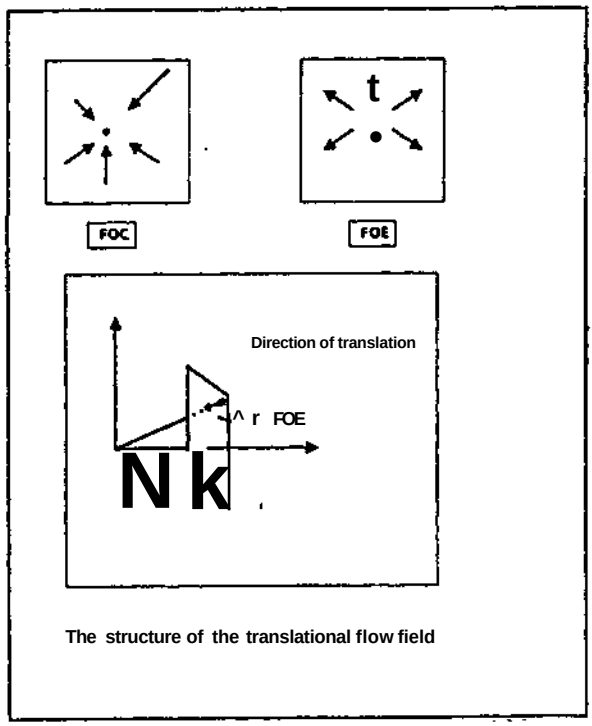


figure 3.3 Focus of Expansion for translational Motion

the limit as A*->0. The image displacement then becomes image velocity, and is called *optical flow*. The optical flow is denoted by the vector (u,v) where:

$$u = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt} \qquad v = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \frac{dy}{dt}$$

Similarly the motion parameters are now the translational velocity *T=(U,V, W)* and the rotational velocity **n** =(<*,£,7) where:

$$T = \lim_{\Delta t \to 0} \frac{\Delta r}{\Delta t} \qquad V = \lim_{\Delta t \to 0} \frac{\Delta r}{\Delta t} \qquad W = \lim_{\Delta t \to 0} \frac{\Delta}{\Delta t}$$

and

$$\alpha = \lim_{\Delta t \to 0} \frac{\omega_x}{\Delta t} \qquad \beta = \lim_{\Delta t \to 0} \frac{\omega_y}{\Delta t} \qquad \gamma = \lim_{\Delta t \to 0} \frac{\omega_z}{\Delta t}$$

Equation (3.6) now becomes:

$$u = (x_0 - x)\frac{W}{Z} + f\beta - \gamma y - \alpha\frac{xy}{f} + \beta\frac{x^2}{f} \qquad (3\text{-}18.1)$$

$$v = (y_0 - y)\frac{W}{Z} - f\alpha + \gamma x - \alpha\frac{y^2}{f} + \beta\frac{xy}{f} \qquad (3\text{-}18.2)$$

where the 3D motion is now characterized by a translational velocity (U,V,W) and a rotational velocity (α,β,γ). Furthermore the FOE is now given by $(x_0,y_0) = (\frac{fU}{W},\frac{fV}{W})$.

Motion perception involves the computation of the parameters of motion from the image displacement field. The latter, becomes in the limiting case, a field of velocities, called optical flow. The relation that optical flow has with the motion parameters, is embodied in equations (3.18). These motion equations involve velocities, both in 3D as well as in the retina. However, in a practical vision system, the retinal measurements that are actually made involve displacements over a small time interval. Thus the above velocity equations, are not strictly applicable, but under certain conditions, the penalty paid for doing this may not be too severe. This happens when the error introduced by the velocity approximation is sufficiently small.

There are two separate approximations embodied in the usage of the equations (3.18) to express the constraints on image motion due to the 3D motion parameters:

(1) *The three dimensional velocity approximation* - The velocity of a point $\rho = (X, Y, Z)$ on a rigid body, moving with a translational velocity $\mathbf{T} = (U, V, W)$, and a rotational velocity $\Omega = (\alpha, \beta, \gamma)$ is given by

$$\frac{d\rho}{dt} = \mathbf{T} + \Omega X \rho$$

Integrating the above with respect to time we have

$$\int_0^{\Delta t} \frac{d\rho}{dt}\, dt = \int_0^{\Delta t} (\mathbf{T} + \Omega\ X\ \rho)\, dt$$

Here X denotes the vector cross product. The three dimensional velocity approximation implies that, for small Δt, the image displacement can be expressed as:

$$\Delta\rho = (\Delta X, \Delta Y, \Delta Z) \simeq \mathbf{T}\Delta t + (\Omega\, \Delta t)\ X\ \rho$$

(2) *The retinal velocity approximation* - This enables us to treat retinal displacements as retinal velocities and is valid so long as $\frac{\Delta Z}{Z} << 1$. This can also be written as relation (A) stated previously.

When both the translational velocity T as well as the depth function Z is multiplied by the same constant, the latter cancels out leaving the equations (3.18) unchanged. The same applies to the equations (3.16). This means that scaling the translation by a constant factor, and at the same time, causing a depth dilation by the same factor leaves the image displacement field unchanged. Thus, from the information available in the image displacement field, the translation vector is only determined up to a scale factor.

In equation (3.16) the depth variable Z is an unknown. An equation relating image displacement to the motion parameters is obtained by eliminating $\frac{t_z}{Z}$ from equations (3.16):

$$\frac{\Delta x(1+\omega_z\frac{y}{f}-\omega_y\frac{x}{f})-(f\omega_y-\omega_z y-\omega_z\frac{xy}{f}+\omega_y\frac{x^2}{f})}{\Delta y(1+\omega_z\frac{y}{f}-\omega_y\frac{x}{f})-(-f\omega_z+\omega_z x-\omega_z\frac{y^2}{f}+\omega_y\frac{xy}{f})}=\frac{x_0-x-\Delta x}{y_0-y-\Delta y}$$

or,

$$\frac{\omega_z\phi_1-\omega_y\phi_2+\omega_z yf+\Delta x}{\omega_z\phi_3-\omega_y\phi_4-\omega_z yf+\Delta y}=\frac{x_0-x-\Delta x}{y_0-y-\Delta y} \tag{3-19}$$

where

$$\phi_1=y\Delta x+xy \qquad \phi_2=f^2+x\Delta x+x^2$$
$$\phi_3=f^2+y\Delta y+y^2 \qquad \phi_4=x\Delta y+xy$$

The above equation relates the motion parameters to the image displacements, which are observables. This is a bilinear equation in the unknown motion parameters. A similar relation is obtained for the differential motion case, by eliminating $\frac{W}{Z}$ from equation (3.18):

$$\frac{u-(f\beta-\gamma y-\alpha\frac{xy}{f}+\beta\frac{x^2}{f})}{v-(-f\alpha+\gamma x-\alpha\frac{y^2}{f}+\beta\frac{xy}{f})}=\frac{x_0-x}{y_0-y} \tag{3-20}$$

In the above analysis, the relations between image motion and 3D motion has been derived by assuming general displacement of a rigid constellation of points in space. This relation is given by equation (3.18). From this, by taking the limiting case, for infinitesimal displacement, the

continuous or differential motion case is obtained. The latter relation can also be obtained directly from the kinematic equations of rigid motion (see Appendix A or [52] for details).

3.5.1. The Information available in the image displacement field

The foregoing analysis illustrates the dependence of the optical flow field on the motion parameters. In other words 3D motion constrains image motion. The magnitude of the translation parameter vector cannot be computed from the optical flow field. The rigid motion parameters observable from monocular retinal optical flow measurements are given by the parameter set $(x_0, y_0, \omega_x, \omega_y, \omega_z)$. Now, we examine the motion equations to see whether the displacement field *uniquely* determines the parameter set.

The question to be answered, before attempting the design of algorithms to compute the motion parameters from optical flow is whether such computation is feasible. This means that given an optical flow field, when can we say that it could be produced by a unique set of motion parameters. The following theorem answers this question, by giving a sufficient condition for uniqueness.

Theorem I: *The optical flow field is uniquely determined by the rigid motion parameters when the moving surface cannot be expressed as a rational function of the form* $\frac{P_1(x,y)}{Q_2(x,y)}$, where P_1 and Q_2 *are polynomials of the first and second*

orders respectively, and (x,y) *are image coordinates.*

Proof: Let a rigid surface Z', moving with translational and rotational velocities (U',V',W') and (α',β',γ') respectively, generate the optical flow field (u,v) given by

$$\begin{aligned} u &= \frac{U' - xW'}{Z'} + f\beta' - \gamma' y - \alpha' \frac{xy}{f} + \beta' \frac{x^2}{f} \\ v &= \frac{V' - yW'}{Z'} - f\alpha' + \gamma' x - \alpha' \frac{y^2}{f} + \beta' \frac{xy}{f} \end{aligned} \tag{3-21}$$

where the translation parameter vector is (U',V',W') and the rotational velocity is (α',β',γ').

Assume that there is another surface $Z(x,y)$ moving with a different set of motion parameters but giving rise to the same optical flow field (u,v), or

$$\begin{aligned} u' &= \frac{U - xW}{Z} + f\beta - \gamma y - \alpha \frac{xy}{f} + \beta \frac{x^2}{f} \\ v' &= \frac{V - yW}{Z} - f\alpha + \gamma x - \alpha \frac{y^2}{f} + \beta \frac{xy}{f} \end{aligned} \tag{3-22}$$

where the 3D motion is now due to a translational velocity (U,V,W) and a rotational velocity (α,β,γ).

From equations (3.21) and (3.22):

$$\frac{U - xW}{Z} \quad \frac{U' - xW'}{Z'} + f\Delta\beta - y\Delta\gamma - \frac{xy}{f}\Delta\alpha + \frac{x^2}{f}\Delta\beta = 0 \tag{3-23.1}$$

$$\frac{V - yW}{Z} - \frac{V' - yW'}{Z'} - f\Delta\alpha + x\Delta\gamma - \frac{y^2}{f}\Delta\alpha + \frac{xy}{f}\Delta\beta = 0 \tag{3-23.2}$$

where, $\Delta\alpha = \alpha - \alpha'$, $\Delta\beta = \beta - \beta'$, and $\Delta\gamma = \gamma - \gamma'$.

Solving for the variable Z' we have: (assuming the focal length f to be unity)

$$Z' = \frac{P_1(x,y)}{Q_2(x,y)}$$

where

$$P_1(x,y) = (UV' - U'V) + x(VW' - V'W) + y(U'W - UW') \qquad (3.24)$$

and,

$$\begin{aligned} Q_2 = {} & (\Delta\beta V + \Delta\alpha U) - x(\Delta\alpha W + \Delta\gamma U) - y(\Delta\beta W + \Delta\gamma V) \\ & - xy(\Delta\alpha V + \Delta\beta U) + X^2(\Delta\beta V + \Delta\gamma W) + y^2(\Delta\alpha U + \Delta\gamma W) \end{aligned} \qquad (3.25)$$

Equations (3.24) and (3.25) imply that the surface Z' that originally generated the optical flow must be a rational function of the form $\frac{P_1}{Q_2}$, to permit ambiguous interpretation of its rigid motion. This is contrary to the the statement of the theorem. This proves the theorem.

Corollary I: *When the motion of a surface is purely rotational, the optical flow field is uniquely determined by the motion.*

Proof: In equation (3.23) make the substitutions $U' = V' = W' = 0$ to obtain:

$$-\frac{U - xW}{Z} + f\Delta\beta - y\Delta\gamma - \frac{xy}{f}\Delta\alpha + \frac{x^2}{f}\Delta\beta = 0$$

$$-\frac{V - yW}{Z} - f\Delta\alpha + x\Delta\gamma - \frac{y^2}{f}\Delta\alpha + \frac{xy}{f}\Delta\beta = 0$$

Now, eliminating Z from the above equations and setting focal length 'f' to unity, we obtain:

$$\begin{aligned} & (\Delta\beta V + \Delta\alpha U) - x(\Delta\alpha W + \Delta\gamma U) - y(\Delta\beta W + \Delta\gamma V) \\ & - xy(\Delta\alpha V + \Delta\beta U) + x^2(\Delta\beta V + \Delta\gamma W) + y^2(\Delta\alpha U + \Delta\gamma W) = 0 \end{aligned}$$

Since this equation must vanish for all values of x and y, the coefficients of unity, x, y, xy, x^2 and y^2 must all vanish, giving rise to the following six equations:

$$\Delta\alpha U + \Delta\beta V = 0$$
$$\Delta\alpha V + \Delta\beta U = 0$$
$$\Delta\beta V + \Delta\gamma W = 0$$
$$\Delta\alpha U + \Delta\gamma W = 0$$
$$\Delta\alpha W + A7J7 = 0$$
$$\Delta\beta W + A7F = 0$$

The above equations imply either $U = V = W = 0$ or Ac* =A£ =A7 =0,

Both these conditions mean that the optical flow field due to a pure rotational motion has a unique interpretation. This proves the corollary.

Corollary II: *It is possible for a flow field generated by pure translational motion to be identical to one generated by another flow field d*. u> both translation and rotation. In other words convergence of the flow vectors directly onto a point on the image plane does not imply purely translational motion.*

The truth of the above corollary will be demonstrated by a numerical example. Consider two flow fields generated by different surfaces undergoing different motions:

In the first case the motion is due to a planar surface given by the equation :

2 6

The motion is rigid and is specified by

7 35

Assume the translation in depth to be unity. Then, from equation (3.18) we have,

$$_{tt}=(z - \frac{7}{-})(i - \frac{1}{-}s+4-y) - 3+5xy - 3x^2$$

$$v = \frac{35}{12} - \frac{139}{36}\left(\frac{35}{6}\right)^2 - \frac{7}{2}xy - \frac{5}{6}$$

In the second case the motion is due to the planar surface given by the equation :

$$Z = \frac{2}{2} - X - \frac{35}{6}Y + 1$$

and the motion is specified by the parameter vector

$$(x_0 = -\frac{1}{2}, y_0 = -\frac{5}{6}, a = 0, p = 0, \gamma = 0)$$

The optical flow field in both the examples are identical.

The question of multiple interpretations of the same flow field, has received some attention in the literature. The foregoing example illustrates the fact that motion of planes can be potentially open to more than one interpretation. It is known (see [81-83, 89]) that the motion of planes have dual interpretations. Uniqueness of interpretation for planes requires three views of four points, or two views of seven points which uniquely define two planes neither of which pass through the origin. In another study Fang and Huang [28] showed that nine points not lying on a second order surface passing through the origin can be used to determine the motion parameters uniquely. Another significant theoretical result is due to Longuet-Higgins [53], and Tsai and Huang [84], where eight points are used to solve for the motion parameters from a set of linear equations. The important question as yet unanswered are, under what conditions the optical flow field is

inherently ambiguous and, what is the degree of the ambiguity possible in optical flow fields. The following analysis answers these questions.

Theorem II. *Under the assumption of rigidity, an optical flow field is amenable to at most three interpretations.*

Proof: Theorem I shows that the optical flow field is enough to determine the rigid motion parameters uniquely for most surfaces. It was seen however that in case of certain rational functions there is potential ambiguity in the interpretation of motion. These are the rational functions of the form

$$Z = \frac{** + by + e \cdot \cdot \cdot}{d + ex + fy + gxy + hx^* + \%y^2} \qquad (3.26)$$

Planar surfaces belong to the above class of surfaces. It has been mentioned previously that planar surfaces can have at most two interpretations. When a surface is non planar, to have multiple interpretations of its motion, it must be of the type given by equation (3.26) with the added property that there is no common factor between the numerator and the denominator.

Let such a surface be undergoing rigid motion (*7',V, W',a,/0,7). Let there be another motion (£/, V, *W,a* + Aa,£ + A#>7 + A7) that produces an identical flow field. Then comparing with equation (3.24) we have

$$\begin{aligned} V'W - VW' &= ak \\ UW - WW &= bk \\ U'V - UV &= ck \end{aligned} \qquad (3.27)$$

where k is some constant factor. Since by definition of the class RJ at least one of a, 6 and e must be non zero, therefore Jfc^O. This is because if k is zero then from the above set of three equations we get the result that the translations (C^V, W[1]) and $\{U,V,W)$ are identical up to a scale factor. Hence by Lemma I of Appendix I, the motion is not ambiguous.

Multiplying the first equation by U', the second by 7^1, and the third by W^1 and adding the three equations we have

$$(aW + 6V'' + cW'yk = 0$$

This means that the motion can only be ambiguous when

$$aU^1 + bV + cW = 0 \quad (3.28)$$

Similarly it can be shown that

$$aJ7 + bV + cW = 0 \quad (3.29)$$

Again comparing the denominator of the rational function in equation (3.26) with equation (3.25), and combining the constant k with the translation parameter $\{U,V,W):$

$$A0V + AaU = d \quad (3.30)$$

$$AaW + A>yU = - t \quad (3.31)$$

$$A0W + A7*^{r} = - / \quad (3.32)$$

$$AaV + ApU = - g \quad (3.33)$$

$$\Delta\beta V + Ar)W = h \quad (3.34)$$

$$\Delta\alpha U + AiW = t \quad (3.35)$$

From equations (3.30), (3.34) and (3.35) we get:

$$2\Delta\alpha U = q \tag{3.36}$$

$$2\Delta\beta V = r \tag{3.37}$$

$$2\Delta\gamma W = s \tag{3.38}$$

where $q = d + i - h$, $r = d - i + h$, $s = -d + i + h$. Substituting from the above equations into equations (3.31), (3.32) and (3.33):

$$qV^2 + rU^2 + 2gUV = 0 \tag{3.39}$$

$$rW^2 + sV^2 + 2fVW = 0 \tag{3.40}$$

$$qU^2 + sW^2 + 2eUW = 0 \tag{3.41}$$

Equations (3.39), (3.40) and (3.41), together with equation (3.28) can admit no more than two solutions. This is because at least one of (q,r,s,e,f,g) must be nonzero. Therefore, since there can be at the most two spurious solutions (recall that the veridical solution corresponds to $k = 0$), the implication is that:

When the optical flow field has more than one interpretation, the number of globally consistent solutions for the motion parameters can be at most three.

This completes the proof of the theorem.

It will be shown that there exist surfaces whose rigid motion induces optical flow that is compatible with *three* distinct interpretations. This fact explains why Longuet-Higgins and Prazdny [52] noted, that from *local* optical flow constraints and their derivatives three interpretations of the

motion are possible since the constraint equations were cubic.

An example of 2D motion field with three distinct rigid motion interpretations:

The equation of the moving surface is

$$Z = \frac{1}{gxy}$$

the motion parameters are $(U',V',0,\alpha,\beta,\gamma)$ the expression for optical flow is therefore

$$u = U'gxy - \alpha xy + \beta(x^2+1) - \gamma y$$
$$v = V'gxy - \alpha(y^2+1) + \beta xy + \gamma x$$

Alternative interpretation I:

$$\frac{1}{Z} = \frac{g}{U}[U'xy - V'(x^2+1)]$$

where the motion parameters are $(U,0,0,\alpha,\beta + gV',\gamma)$. The optical flow field is given by

$$u_1 = U\frac{g}{U}[U'xy - V'(x^2+1)] - \alpha xy + (\beta + gV')(x^2+1) - \gamma y$$
$$v_1 = -\alpha(y^2+1) + (\beta + gV')xy + \gamma x$$

Alternative interpretation II:

$$\frac{1}{Z} = \frac{g}{V}[V'xy - U'(y^2+1)]$$

The motion parameters are $(0,V,0,\alpha - gU',\beta,\gamma)$. The optical flow field is

$$u_2 = -(\alpha - gU')xy + \beta(x^2+1) - \gamma y$$
$$v_2 = V\frac{g}{V}[V'xy - U'(y^2+1)] - (\alpha - gU')(y^2+1) + \beta xy + \gamma x$$

It is easily verified that $u = u_1 = u_2$ and $v = v_1 = v_2$.

Theorem I states that under certain cases the optical flow field may not indicate the motion parameters uniquely. The next theorem shows how unambiguous determination of the motion parameters can be achieved from optical flow data.

Theorem III: ***The ambiguity of the optical flow field disappears when the observation period extends over more than two time instants, assuming that the motion in three space is steady.***

Proof: The term steady motion indicates that the direction of translation is fixed in space with respect to any inertial frame of reference. In other words, the observer's line of trajectory is a straight line.

The proof of the theorem follows straightforwardly from equations (3.28) and (3.29). Those equations tell us that ambiguity can only occur when the direction of translation lies on the plane tangent to the observed surface at the origin. Since this condition must necessarily be maintained, in order to preserve ambiguity, we can state:

To maintain ambiguity, the spatial trajectory of the observer's nodal point (i.e. origin of the frame of observation) must lie on the observed surface.

Since the observer's trajectory is a straight line, the above condition implies:

(a) The surface is planar.

(b) It is developable, i.e. one of the principal curvatures vanishes at all points (e.g. a cylindrical surface).

In the first case it can be shown that the ambiguity cannot be sustained [83]. In the second case, the direction of translation must be along the principal axis corresponding to the vanishing principal curvature. This means that all the interpretations must have their translational velocities in the same direction. Thus their rotational velocities must be identical (see appendix A). Hence the motion will not have ambiguity. This completes the proof of the theorem.

Another way of resolving the ambiguity in the optical flow is by using shape information. There is a strong relationship between the parameters of motion, the optical flow field and the structure of a moving surface. The structure of the surface is defined by depth ratios between any pair of given points (see Appendix B). The following propositions makes this concept clear.

Proposition I. *When the parameters* (i.e. $x_0, y_0, \alpha, \beta, \gamma$) *describing the motion of a rigid surface are known then the structure of the surface is uniquely determined from the optical flow field.*

Proof: The proof is evident from equation (3.18). Note that we can obtain the depth function up to a constant dilation factor W. In other words the ratio of depths at any two image points can be computed.

Proposition II. *When the structure of a surface is known then the parameters describing its rigid motion are uniquely obtained from the optical flow generated*

by the motion.

Proof: See Appendix IL

Even the partial specification of shape can lead to a correct perception of rigid motion. A illustration of the fact that shape information can disambiguate between alternative motion interpretations comes from the next theorem.

Theorem IV: ***The motion of a planar surface whose direction of translation does not lie in the plane of its surface normal and the line of sight, can be interpreted correctly from the optical flow generated, when the tUt of the plane is known.***

Proof: Let the equation of the planar surface be

$$z = \frac{d}{1 - px - qy}$$

where (p,q) is the orientation of the depth plane and 'd' is the distance from the origin along the z axis (e.g. line of sight). Substituting the above into equation (3.18) and observing that we can ignore multiplication of the translational parameters by a constant (such as d) since we can compute the former up to a scale factor anyway, we have:

$$\begin{aligned} tt &= l_x - l_2 x - l^{\wedge} y + U x y + l_B x^2 \\ v &= /_{\text{„}} - l_7 x - l_B y - l_h x y + /_4 y^2 \end{aligned} \qquad \textbf{(3.42)}$$

where the unknowns $\{ a_f \}$ are given by

$$U + p^{\wedge} li \qquad (3.43.1)$$

$$Up + W = l_2 \quad (3.43.2)$$

$$Uq + 7 = /s \quad (3.43.3)$$

$$Wq - a = i_4 \quad (3\text{-}.43.4)$$

$$Wp + fi = l_h \quad (3.43.5)$$

$$V - a = /_e \quad (3.43.6)$$

$$7 - Vp = ^ \quad (3.43.7)$$

$$V? + W = /_{„} \quad (3.43.8)$$

If we can estimate the *synthetic parameters* {'. ^{y} b" .^ung measurements of optical flow at a minimum of four aaiuiuie points, in the image and, *in addition* can measure the *tilt of* the depth plane, i.e.

$$| - r \quad (3.44)$$

Then from (3-43.7) and (3-43.8) and (3-44) we have:

$$7 + rW = /_T + r/_t \quad (3.45.1)$$

From (3-43.2), (3-43.3) and (3-44) we have :

$$r_7 \_ w = r/_s - l_2 \quad (3.45.2)$$

Therefore, since r^2 + 1^0 we have:

$$\gamma = \frac{\text{illegible}}{\text{illegible}} \quad (3.45.3.1)$$

$$W = \frac{l_{\cdot} - '(', - ',) + ^'.}{T^2 + 1} \quad (3.45.3.2)$$

Now if $W \wedge l_9$ (i.e. ? 7^ 0) we have from (3-43.8) and (3-43.3):

$$\frac{Uq}{vq} \_ \frac{U}{V} \_ \frac{l_3 - \gamma}{W} \_ \quad (3454)$$

(iAbA)

otherwise if l_7 5^ 0 (i.e. p ^ 0) we have from equations (3-43.7) and (3-43.2):

$$\frac{Up}{Vp} = \frac{U}{V} = \frac{l_2 - W}{\gamma - l_7} = k \qquad (3\text{-}.45.4)$$

(if both p and q are zero then the parameters are easily solved for)

Now from (3-45.4), (3-43.6) and (3-43.1) we have:

$$k\alpha + \beta = 1 - kl_6 \qquad (3.45.5)$$

Also from (3-43.5) and (3-43.4) we have:

$$\tau\alpha + \beta = l_5 - \tau l_4 \qquad (3.45.6)$$

Therefore, since $\tau \neq k$, from the assumption made in the statement of the theorem, then equations (3-45.5) and (3-45.6) are independent, and we have:

$$\alpha = \frac{(1 - kl_6) - (l_5 - \tau l_4)}{k - \tau} \qquad (3.45.7.1)$$

$$\beta = \frac{k(l_5 - \tau l_4) - \tau(1 - kl_6)}{k - \tau} \qquad (3.45.7.2)$$

Now U and V can be determined from equations (3-43.6) and (3-43.1). Thus we have determined the motion parameters uniquely from the optical flow and tilt information.

At this point it may be mentioned in passing that it is possible to obtain the motion parameters uniquely from the optical flow generated by two planes moving together rigidly. In this case the optical flow is locally second order. If the eight synthetic parameters are now measured at two different regions of the flow field then

$$
\begin{aligned}
U\Delta\frac{1}{d} &= \Delta l_1 \\
U\Delta\frac{p}{d} + W\Delta\frac{1}{d} &= \Delta l_2 \\
U\Delta\frac{q}{d} &= \Delta l_3 \\
W\Delta\frac{q}{d} &= \Delta l_4 \\
W\Delta\frac{p}{d} &= \Delta l_5 \\
V\Delta\frac{1}{d} &= \Delta l_6 \\
-V\Delta\frac{p}{d} &= \Delta l_7 \\
V\Delta\frac{q}{d} + W\Delta\frac{1}{d} &= \Delta l_8
\end{aligned}
\tag{3.46}
$$

where the two planes involved in the motion are given by $z = \frac{d}{px + qy + 1}$ and $z = \frac{d'}{p'x + q'y + 1}$. The Δ operator in front of any quantity denotes the difference of the corresponding parameters for the two planes, e.g. $\Delta\frac{p}{d} = \frac{p}{d} - \frac{p'}{d'}$.

The above equations imply that when at least one of, $\Delta\frac{p}{d}$ or $\Delta\frac{q}{d}$ or $\Delta\frac{1}{d}$ is non zero the translational parameters are uniquely determined. Hence in such a case the rigid motion parameters are determined uniquely from the optical flow field (see Appendix A). Therefore

When two planes, neither of which pass through the origin, move rigidly together, their motion is uniquely determinable from the optical flow field generated.

3.5.2. Summary of the perspective projection case

The analysis presented here leads to considerable insight into the 3D motion interpretation problem. Previous results (e.g. [28, 84]) by Huang and his colleagues presented *sufficient* conditions for uniqueness of three dimensional motion interpretation, since, they were concerned with specific algorithms. The work, reported here, deals with necessary conditions for unique interpretation of 3D motion from the optical flow field.

While the surface denoted by equation (3.26) does mean second order surfaces containing the nodal point of the camera, it is certainly true that all such surfaces do not admit ambiguous interpretations of their 3D motions. Multiple interpretations require, in addition, that the the constraints given by (3.29), (3.39), (3.40) and (3.41) all be satisfied.

Thus consider, an algorithm, such as Prazdny's [67], where nonlinear (and independent) flow constraints at five retinal locations are used to obtain a 3D motion interpretation. It is now possible to answer the question as to whether the solution obtained is the only one possible. Since now a set of motion parameters is known, from equation (3.22) the relative depth $\frac{Z}{W}$ can be obtained at the five retinal locations. The latter, when substituted into equation (3.26), generates five linear equations in the surface parameters a,b,c,d,e,f,g,h,i. These together with the four constraints (3.29), (3.39), (3.40) and (3.41) constitute nine linear

homogeneous equations in the nine surface parameters. Therefore uniqueness of interpretation is possible if the determinant of the above system is zero. Which in turn implies, that all the surface parameters must be zero. This makes it impossible to construct any other interpretation from measurements at the five retinal locations, guaranteeing that the solution obtained is the only one possible.

3.6. Summary of motion constraint results

Uniqueness proofs of the type derived by Tsai and Huang and Fang and Huang do not allow us to visualize the situations when the optical flow field is intrinsically ambiguous, admitting more than one interpretation. The analysis of the optical flow field to determine cases of ambiguity was a major focus of this chapter. We saw that three temporally contiguous image frames contain enough information to uniquely recover 3-D motion and structure under perspective projection. Since the optical flow field (two temporally proximal frames) is, in general, ambiguous, two frames can recover structure when the moving surface satisfies the conditions of Theorem I.

The image formation geometry used in the analysis involved the perspective projection. We also briefly examined an approximation of the above model called orthographic or parallel projection. The attendant simplicity in the motion constraint equations can be used to considerable

advantage in the preliminary analysis of the motion perception problem.

The following results were derived:

1. The component of rotation about the line of sight, the ratio of the other two components of rotational velocity, and the tilt function is uniquely computable from a single optical flow field, for a rigid *non planar* surface.

2. When the surface normals for a rigid surface are known then the motion parameters can be computed uniquely.

The *Perspective Projection model* (see figure 3.2) is a more accurate model of image formation by eye or camera. For this model it is proved that:

1. The optical flow field, under the assumptions of rigidity can have at most *three* interpretations.

2. The rigid motion of any surface whose depth from the nodal point of the sensor cannot be expressed by the rational function $\frac{P_1(x,y)}{Q_2(x,y)}$, where P_1 and Q_2 are rational functions of the first and second orders respectively, is uniquely computable from the information in the optical flow field.

3. Two optical flow fields, obtained at different time instants, determine the motion parameters uniquely.

4. The motion parameters are uniquely determined from the optical flow field when the corresponding motion involves rotation only.

5. The optical flow due to planar surfaces is generally ambiguous. However this ambiguity can be resolved either when the flow field is due to more than one plane moving together rigidly, or in the case of a single plane, if its tilt is known.

Chapter Four

Algorithms for Rigid Motion Perception

4.1. Introduction

In this chapter the applicability of the *Hough Transform* technique to motion parameter estimation is examined experimentally.

The main difficulty in computing the 3D Rigid Motion parameters is that the equation constraining the image motion to the 3D motion is nonlinear. Another complication arises from the high dimensionality of the parameter space. If it were possible to separate the component of the image displacement due to translation from that due to rotation we could have efficient algorithms for the computation of the three dimensional motion parameters.

The brute force hough algorithm [7] is seen to have limitations that stem from the above difficulties. The next section of this chapter outlines the various computational strategies that were adopted to overcome the above problems. After this, algorithms employing these strategies and their

experimentally determined performance is presented. The remaining portion of this introductory section is a brief discussion of other algorithms that have been proposed in the literature.

An algorithm due to Ullman uses a simplified situation where the rotation axis is assumed to be along the z axis [85] • The constraint he obtained was an equation of the fourth degree in the sine of the rotation angle. Roach and Aggarwal derived a set of nonlinear constraint equations in eighteen parameters to characterize rigid body motion [71]. In recent years, most of the work in motion interpretation in the literature attempt either least square error minimization or iterative search techniques to compute the set of motion parameters that best describe the image motion data. The constraint equation used is some form similar to the one derived in chapter three (also equation 4.7). Brass and Horn [19] compute parameter set that minimizes the square of the error between the measured optical flow and that computed from the parameter constraint In general such a technique will give rise to a system of non-linear equations from which the parameters must be computed using some suitable iteration scheme. Longuet-Higgins and Prazdny [52] mention the possibility of using motion parallax to simplify the computation of the global motion parameters. Lawton and Rieger [70] uses a similar idea to factor out the rotational component of the optical flow at depth discontinuities or regions where the depth gradient is large. This method is not reliable since it hinges

upon the ability to compute flow vectors reasonably accurately at discontinuities. Since almost all algorithms, to date, for computing optical flow face problems at regions where the field is sharply discontinuous.

Another method is to linearize the constraint equation by writing equation 4.7 as a linear equation in *eight* parameters* Obviously these eight parameters are each functions of the values of the five actual parameters. This implies that linear least square methods are not applicable here, since the eight *synthetic* parameters are not independent of one another. A similar method is used by Tsai and Huang [84] but they found that the computation is very sensitive to errors. Algorithm V attempts to alleviate the problem of high dimensionality of the above scheme by using spatio-temporal derivatives of the optical flow field.

In the case of *General Motion,* where one or more objects move with respect to the observer, the situation is complicated by the fact that we have to determine several sets of parameters, corresponding to the several bodies in motion. However, the image motion measurement technique of chapter two has been found to be quite good under such circumstances. This fact enables us to assume that the algorithm for motion parameter estimation can deal, without loss of generality, with the motion of a single body. Motion segmentation has also been studied in restricted domains by Fennema & Thompson [33]. The more tricky question concerning the difference between *Egomotion* and *General Motion* has to do with the choice

of the body frame of reference. In the case of egomotion the camera frame and the natural body frame can be though to coincide. This means that the notion of steady motion (i.e. parameters relatively unvarying within any small period of observation) is intuitive, in the sense that it implies steady motion of the observer in space. On the other hand for the motion of an object the usual convention is to chose the body frame of reference to coincide with the camera frame. In this case the steady motion of the body in space need not imply steadiness of the observed motion parameter values.

Recently a way of determining motion parameters from three dimensional flow has been suggested [8]. This method is amenable to adaptation to the general motion case. It is not clear as to how difficult it is to compute the three dimensional flow in this case. However, it can be shown that in case a depth map can be obtained (by some stereo matching technique), the three dimensional flow map can be calculated.

Computer algorithms for determining the parameters of rigid motion are discussed in the light of the various constraints developed in the previous chapter. The treatment will consider both *orthographic* and *perspective* projections. Some of the algorithms are described in detail, while others are outlined, particularly when they bear any similarity to one already discussed. In the algorithms proposed here, the *Hough Transform* technique [7] is used to compute the desired global parameters from sets of constraint

equations obtained at different image (or retinal) locations. It should be noted that least square error minimization techniques are also applicable in most cases.

Recall that the notation for optical flow is (u,v). While the translation parameters are denoted by (U,V,W) or $(x_0 = \frac{U}{W}, y_0 = \frac{V}{W})$ and the rotational parameters by (α,β,γ).

4.2. Using the Hough Transform for Motion Parameter estimation

The concept of the hough transform is very simple. It is closely related to the idea of clustering, introduced in chapter two. Consider an example problem where we are required to estimate the parameters of a straight line in two space from local measurements of small edge segments. The form of the line equation we will use is

$$x \cos\theta + y \sin\theta = \rho$$

and hence the parameters to be estimated are (ρ,θ). The set of measurements is given by

$$M = \{(x_i,y_i,\theta_i) \mid \textit{there is an edge at } (x_i,y_i) \textit{ with orientation } \theta\}$$

Using the elements of M we obtain a distribution $H(\rho,\theta)$ which denotes the count of the number of times each of the (ρ,θ) pairs satisfied the line constraint equation for all the data values. This distribution is called the hough transform of the data set M. The parameter estimate (ρ^*,θ^*) is then given by the mode of the distribution $H()$. The situation is depicted in

figure 4.1.

Hough Transform is defined in the (ρ,θ) parameter space. An important aspect of the method is the necessity of quantizing (or discretizing) the parameter space in order to implement the transform process by computer (or by a hardwired connectionist network [31]). The degree of quantization is, in most cases, a critical control variable. The quantization can be visualized by imagining the parameter space to be covered by a set of cells that collect evidence or votes from the data values in order to determine whether the desired parameter set lies in the space spanned by the cell.

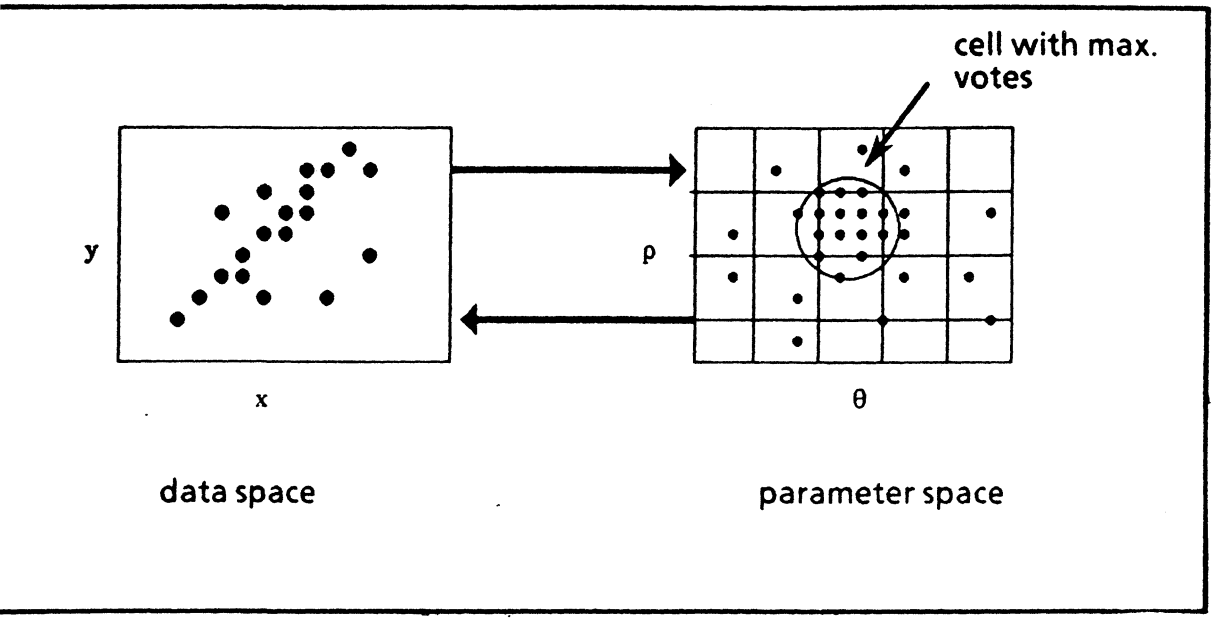


Figure 4.1 Parameter estimation by Hough Transform

An alternative formulation arises when we are unable to measure the orientation θ of the edge elements. In this case

$$M = \{(x_i, y_i) \mid \textit{there is an edge at } (x_i, y_i)\}$$

Therefore, every (x_i, y_i) determines a constraint surface in the parameter space :

$$x_i \cos\theta + y_i \sin\theta = \rho$$

Thus the transform is obtained by voting for every cell in the transform space that "satisfies" the constraint for a given data element. Again the estimated parameter set (ρ^*, θ^*) is obtained by taking the mode of the resulting distribution $H(\rho, \theta)$.

The motion parameters that are to be estimated are:

$$(x_0 , y_0 , \alpha , \beta , \gamma)$$

The measured data is the optical flow field $[u(x,y), v(x,y)]$. In order to use the hough transform method for to tackle the motion perception, the following issues have to be addressed:

(i) What does it mean for the data to satisfy a constraint? This question is important since we have to contend with nonlinearity, discretization and noise. Thus the data may never exactly satisfy the constraint.

(ii) At what coarseness level should the parameter space be quantized.

(iii) How does nonlinearity affect the first two issues.

In order to represent the the parameter space one has to be able to determine the bounds of the plausible parameter values. This does not seem an unreasonable demand, however a more critical factor is the quantization of the space. In of nonlinear constraints small discretization errors may cause large fluctuations in the constraint surface. Hence with coarse discretization the issue of "constraint satisfaction" may be difficult to determine. This is reflected by the results obtained from algorithm III.

A heuristic solution to this problem is to stipulate that constraint satisfaction implies that the constraint surface intersects the parameter cell in question. This leads to a simple scheme to determine intersection, in the case of linear surfaces, whose distance from the cell center can be determined by substituting the cell center parameter values in the constraint equation. This is however not possible in the case of nonlinear constraint surfaces (figure 4.2).

The great advantage of this method is that very coarse quantization of the parameter space is possible. The only problem with this intersection strategy is that when the cells are large the distribution obtained may be multimodal. This situation is depicted in figure 4.3. In this case the spurious modes have to be eliminated by successive refinement by considering particular candidate cells and splitting them into sub cells and repeating the voting process. This strategy is used in algorithm V and a modified version of algorithm III.

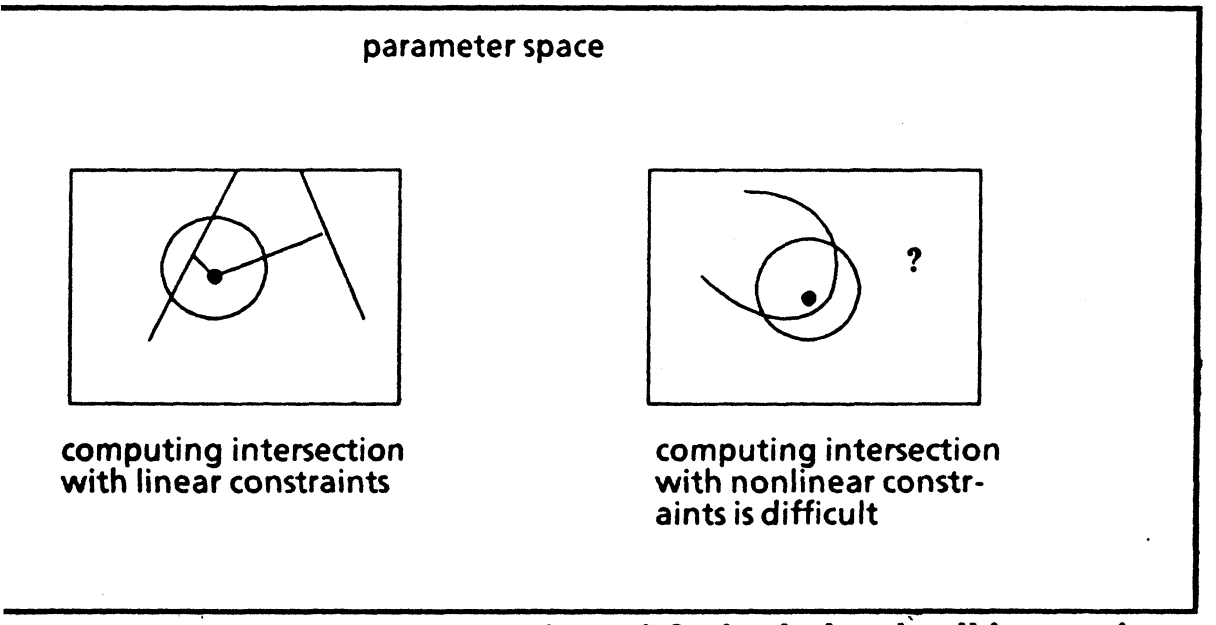


Figure 4.2 Determining Constraint satisfaction by hough cell intersection

In some of the subsequent algorithms the five dimensional parameter space is subdivided into a translational subspace and a rotational subspace. The first subspace is quantized in terms of rectangular cells, while in the case of the rotational space we have used a gaussian sphere representation, using geodesic tessellations, to span the directions in three space corresponding to the axis of rotation.

The results reported in this chapter indicate that hough transform can be a reliable and robust method for motion parameter estimation. The problem of nonlinearity cannot be totally be removed, necessitating the knowledge of some initial estimate of the parameter solution set. without this it becomes difficult to label the parameter cells in the transform space and requiring a large number of such cells.

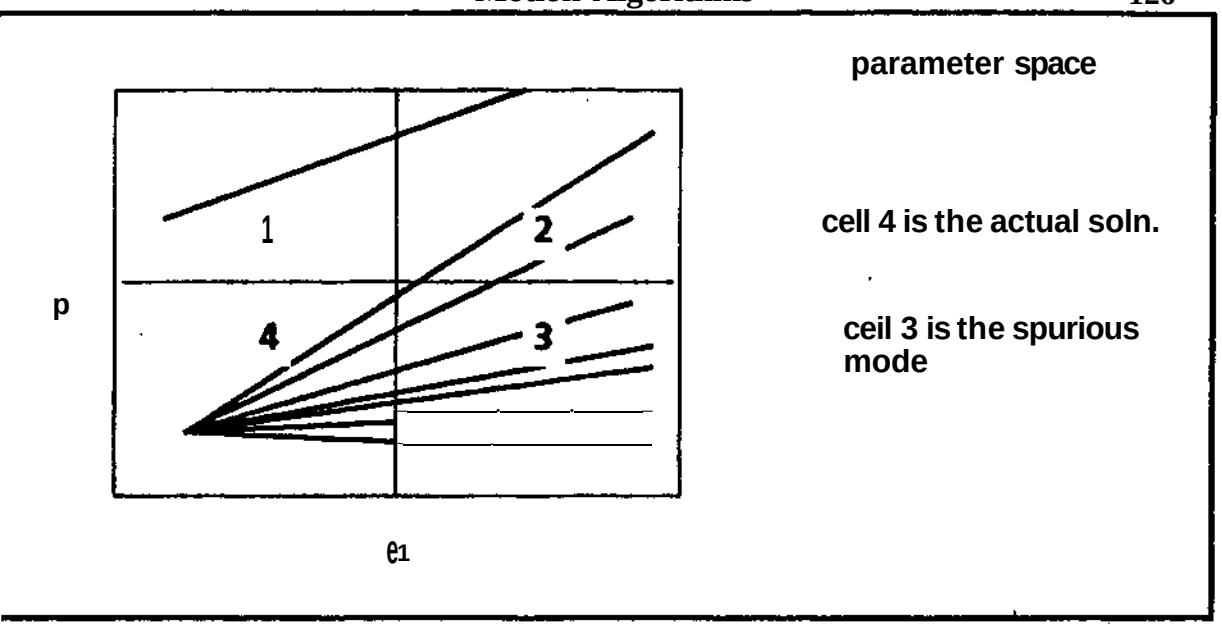


Figure 4.3 Spurious mode formation in the cell intersection scheme!

4.3. Motion under Orthography

For the sake of completeness the case of orthographic projection is considered in this section. This is , however a restrictive situation, which is approximates the imaging geometry when the imaged object is either very far away, or the focal length of the lens is large. This case has been analyzed extensively in the literature, [5, 42, 79], are some examples. The above methods deal with local analysis of the image motion field. The algorithms presented in this section are based on the uniqueness results of chapter three and involve global analysis of the optical flow field.

Under orthography the translational part of the optical flow field is constant and hence the translational parameters are not computable. Hence motion parameters here, always refer to the rotational velocity parameters (α,β,γ).

The relevant equations are

$$\underset{A}{Au} = \underset{A}{\pounds A2r} - \underset{A}{7Ay} \qquad (4\text{«}1)$$

where the A symbol denotes that the following quantity is a difference obtained from measurements made at two different retinal locations. The relation between the surface gradients and the optical flow derivatives are:

$$Ji_{z} \_,\pounds \qquad (4.2.1)$$

$$\pounds \text{——} \pounds^{+}7 \qquad (4.2.2)$$

$$4^{\wedge} = /j - |^{\wedge} - f \qquad (4.2.3)$$
$$ay \qquad ay$$

$$\frac{*!!}{\partial y} = \_ \; {}^{a}i^{*}\frac{fL}{\partial y} \qquad (4.2.4)$$

Algorithm I: *Motion parameters from image motion and structure information.* The simplest instance is when the structure of the moving object is known. In the discrete case the relative depth function, AZ(x,y), values are enough to compute the parameters (<*,0,7) uniquely from the linear equation (4.1). For the differential case structure or shape can be represented by the surface normals $(-\frac{\partial Z}{\partial x}, -\frac{\partial Z}{\partial y})$. If the surface normals are known everywhere, then we can integrate the surface normals to obtain the depth up to a constant additive term. In other words *AZ(x,y)* is computable. In this case measurement of optical at three non collinear points is enough to compute the rotational parameters. However, if the surface normals are only known at sparse locations, but the optical flow field is locally known at these

locations then we can use equation (4.2) for computing the rotation parameters. In this case we are relying on the fact that the first derivatives of the flow can be reliably computed. This is possible when, in the neighborhood of the points of interest, the optical flow values have been measured at enough locations so as to allow analytic reconstruction of the optical flow function. Finally note that, if the motion parameters are known then the structure can be obtained from the image motion for both the discrete and the differential cases. The steps in the algorithm are:

1. Set up a three dimensional accumulator array for the rotation parameters: $h[\alpha,\beta,\gamma]:=0$.

2. For every point in the image where optical flow *and* surface normals are known, select the constraint equation (4.1) if the estimated measurement error in the surface normal function is less than that estimated for the optical flow function; otherwise select equation (4.2).

   For all values of (α,β,γ):

   If (α,β,γ) satisfies the constraint equation selected

   $h[\alpha,\beta,\gamma]:=h[\alpha,\beta,\gamma] + 1$

3. Obtain the maximum value in the accumulator array. The corresponding indices are the desired values for the rotation parameters.

Algorithm II; ***Motion parameters and structure from image motion.*** **When the structure is not known then, considering the differential case and eliminating** $(\frac{\partial Z}{\partial x}, \frac{\partial Z}{\partial y})$ **from equations (4.2) :**

$$\frac{dv}{\partial x} = -\frac{a}{\beta}\frac{du}{\partial x} + \gamma \qquad (4.3.1)$$

$$\frac{\partial u}{\partial y} = -\frac{\beta}{a}\frac{\partial v}{\partial y} - \gamma \qquad (4.3.2)$$

Similarly, eliminating AZ from equation (4.1):

$$fiu - 7X + p7y + v = 0 \qquad (4.4)$$

where $p = -\frac{\S}{p}$.

It is easy to obtain quadratic equations in either 7 or $\frac{\alpha}{p}$ **from the equations (4.3). This means that in general, at every image location, from the measurement of the spatial derivatives of the optical flow at most two sets of values of the parameters** $(\frac{\alpha}{p}, 7, \frac{\beta}{9})$ **may be obtained. However, if some global assimilation technique, like the hough transform (see [7]) is used, then, as shown previously, if the moving surface is non planar, only one set of parameters will be globally consistent. An exactly similar method, but using differences of image displacements, can be devised for the discrete case starting from equation (4.4).**

4.4. Motion under Perspective projection

The relation between the optical flow and the motion parameters is given by the equation:

$$u = \frac{U - xW}{Z} - \alpha xy + \beta(x^2 + 1) - \gamma y$$
$$\frac{}{Z} \qquad \cdot \beta xy + \gamma x \tag{4.5}$$

From the above we obtain, by eliminating Z:

$$\frac{u - f\,axy - /3(x^2 + 1) + 7y}{t + a(f^* + 1).\ \wedge + 7^*} = \frac{U - xW}{\wedge - yW} \tag{4.6}$$

Observe from the right hand side of the above equation, that its value is unchanged when the translational parameters are multiplied by some constant. Hence we can determine the translational parameters only up to a scale factor. If we assume that $W \wedge 0$ then the previous equation can be written as:

$$\frac{u + axy - P(x^2 + 1) + 7F}{v + a\{y^2 + 1) - fixy + 7^*} = \frac{x_0 \sim *}{y_0 - y} \tag{4.7}$$

If $w = 0$ then (4.6) reduces to:

$$\frac{tt + otsy - £(s^2 + 1) + iy}{* + a(*^2 + l) - j8*y + 7^*} = \frac{U}{\wedge} \tag{4.8}$$

Equations (4.6), (4.7) and (4.8) are bilinear in the translation and the rotation parameters. This nonlinearity makes it difficult to combine constraints from different image locations to compute the motion parameters. To summarize, the problems with computation of motion parameters are:

1. The constraint equations are nonlinear.

2. The parameter space is of high (e.g. five) dimensionality.

Algorithm III: *Hough transform in 5D parameter space.* This type of algorithm can be easily realized by simple parallel neuronal hardware (see [31]). The parameters that are to be determined are the polar angles (or direction cosines) representing the directions of translation and rotation, and the magnitude of the rotation vector. This representation for the rigid motion parameters is convenient since the parameter subspaces representing directions in space become easy to quantize by means such as *geodesic tessellation* of the gaussian sphere. The steps in the algorithm are:

1. Select a coarseness scale for the parameter subspaces. For instance, how many distinct directions in space, the range of values estimated for the rotation magnitude and the sampling interval in this range. Initialize the parameter units belonging to the *hough transform space* (this is the five dimensional accumulator array where the *votes* for every parameter vector is tallied).
2. For all retinal locations where optical flow has been measured do step 3:
3. For all possible parameter values (i.e. values of the parameter quintuple) admitted in step 1, do:

   (i) If the direction of the translational velocity is not parallel to the image plane select equation (4.7) else select equation (4.8).

   (ii) If the parameter values satisfy the chosen constraint equation vote for the corresponding parameter vector.

4. Find the parameter quintuple that has received the maximum number of votes.

5. Restrict the parameter space to a neighborhood of the selected parameter quintuple. Repeat the steps from 2 to 4 after choosing a finer parameter space quantization.

6. If the error due to the parameter quantization is acceptable then stop and return the parameter values computed. Otherwise repeat step 5.

Some Remarks:

(i) The space and time required by the algorithm is reduced by periodically examining the parameter accumulator units and purging those that have collected only a few votes compared to the top contenders. This is possible, since it is assumed that the noise in the optical flow data is uniformly distributed in retinal space.

(ii) The confidence of the computed parameter quintuple is the ratio of the votes it received to the maximum votes possible.

(iii) If in step 4 instead of a clear winner, a number of contenders are found then step 5 might have to be repeated for each of these for finer resolutions. Then the winner is the parameter quintuple that comes through with the highest confidence.

Algorithm III, performs well when the quantization of the parameter space is not too coarse. The following table 4.1 shows the degradation of

performance with coarser quantization. Although the motion constraint equation is nonlinear it is actually bilinear in form. This means that if either the two translation parameters are known then the constraint becomes linear in the three rotational parameters. The same holds true when the three rotational parameters are known in which case the constraint becomes linear in terms of the two translational parameters. This fact is used to modify Algorithm III, so that instead of voting in a five dimensional space, we break up the parameter space into two subspaces corresponding to the translational and rotational parameters respectively. The two subspaces are arranged in a hierarchical fashion. Every cell in the translation space spawns a rotational space where the linear intersection strategy is used to accumulate votes. The method is depicted in figure 4.4. It was found that this strategy was very robust with respect to parameter space quantization. In fact very coarse translation as well as rotational spaces could be used and successively refined. Plate 4.1 shows some of the results for the rotational parameter. The displays show vote distribution on the geodesic gaussian sphere which has been quantized at various resolutions. The quantization parameter N denotes the number of distinct directions in space used for the

Quantization Error		Computed Parameters					Error	
Trans.(%)	Rot. (%)	x_0	y_0	α	β	γ	Trans.(%)	Rot.(%)
10.0	8	0.44	1.4	3.0	2.5	2.7	30	20.0
5.0	4	0.97	2.4	3.3	2.7	4.0	9	5.0
2.5	2	1.00	2.0	3.0	2.0	4.4	6	0.5
1.2	1	0.99	1.9	3.0	2.1	3.9	3	1.0

Table 4.1 Quantization effects on five parameter hough transform

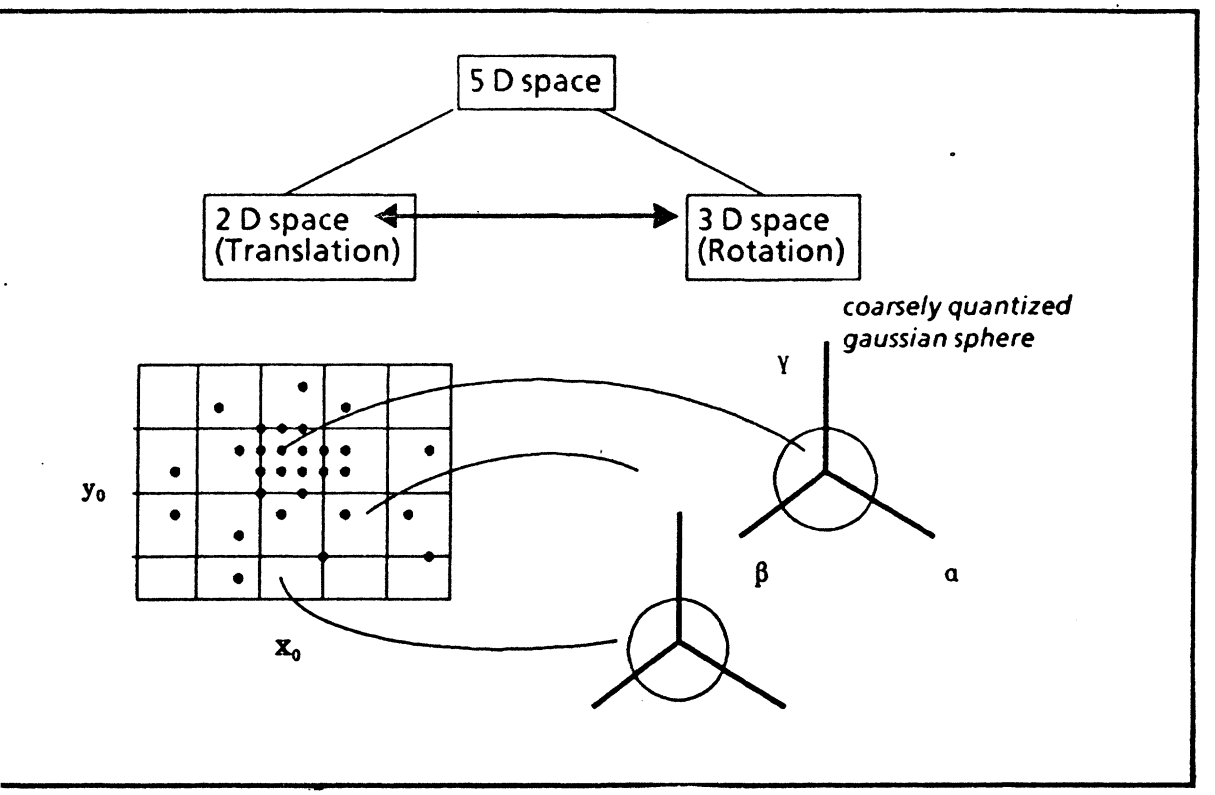


Figure 4.4 Hough Transform with Decoupled Subspaces

The next two algorithms approach the nonlinearity problem by linearizing the constraint equation. Although in this case the price we pay is that the dimensionality of the parameter space increases. In the following discussion it is assumed that the not all the translational velocity components are zero. This is a valid assumption since it has been shown in a previous section that the motion parameters for pure rotational motion are uniquely detectable.

From equation (4.6) we have:

$$(yu - xv)W + vU - uV - x(\alpha W + \gamma U) - y(\beta W + \gamma V) - xy(\alpha V + \beta U) + x^2(\beta V + \gamma W) + y^2(\alpha U + \gamma W) \quad (4.9)$$

Now we state and prove a lemma regarding the feasibility of computing the

motion parameters using the constraint given above.

Lemma I: *The optical flow components can be expressed as an implicit polynomial equation* F(tt,9,x,y;pj9»' = 1,. . 8) =0 *involving the image coordinates (x,y) and eight* **linearly independent** *parameters p_4 unless the depth function is a rational function* $\frac{P_l(z,\ y)}{Q^{x}\!\left(\ ,\ \right)}$ *where P_x and Q_2 are polynomials of first and second ordersrespectively.*

Proof: Equation (4.9) is homogeneous in the motion parameters. Assume that the parameter $W \wedge 0$ (The case where $W = 0$ but either U o r V ^ O **can be worked out in an analogous manner). Dividing the above equation by W yields:**

$$\{yu - xv) + p_t v - p_2 u \text{ 4- } p_s - \text{P4*} - PbV + \text{P**}^2 + PiV* - Vt*V = 0 \qquad (4.10)$$

where

$$Pi = x_0 \qquad (4.11a)$$

$$\text{P2-yo} \qquad (4.11b)$$

$$p_3 = \alpha x_0 + \beta y_0 \qquad (4.11c)$$

$$p_4 = \alpha + \gamma x_0 \qquad (4.11d)$$

$$\text{Ps} = 0 + 7*0 \qquad (4.11e)$$

$$p_6 = \gamma + \beta y_0 \qquad (4.11f)$$

$$p_7 = 7 + a z_0 \qquad (4.11g)$$

$$p_8 = \beta x_0 + \alpha y_0 \qquad (4.11h)$$

The parameters p,'s are linearly dependent if and only if

$$k_1 v - k_2 u + k_3 - k_4 x - k_5 y + k_6 x^2 + k_7 y^2 - k_8 xy = 0 \qquad (4.12)$$

where the k_i's are constants not all of which are zero. Let the optical flow be due to a rigid surface Z moving with velocity $(\overline{U}, \overline{V}, \overline{W}, \overline{\alpha}, \overline{\beta}, \overline{\gamma})$. In this case:

$$\begin{aligned} u &= \frac{\overline{U} - x\overline{W}}{Z} - \overline{\alpha}xy + \overline{\beta}(x^2 + 1) - \overline{\gamma}y \\ v &= \frac{\overline{V} - y\overline{W}}{Z} - \overline{\alpha}(y^2 + 1) + \overline{\beta}xy + \overline{\gamma}x \end{aligned} \qquad (4.13)$$

Assume that the parameters p_i are linearly dependent. This implies that in equation (4.12) there must be at least one k_i that is not equal to zero. However, if both k_1 and k_2 are zero, then, all the k_i's must be zero. Hence, if the parameters p_i are linearly dependent, then at least one of k_1 and k_2 must be nonzero.

Substituting for 'u' and 'v' in equation (4.12) from equation (4.13) we obtain:

$$k_1 \left(\frac{\overline{U} - x\overline{W}}{Z} - \overline{\alpha}xy + \overline{\beta}(x^2 + 1) - \overline{\gamma}y \right) - k_2 \left(\frac{\overline{V} - y\overline{W}}{Z} - \overline{\alpha}(y^2 + 1) + \overline{\beta}xy + \overline{\gamma}x \right)$$
$$+ k_3 - k_4 x - k_5 y + k_6 x^2 + k_7 y^2 - k_8 xy = 0$$

Since both k_1 and k_2 are not zero, we obtain Z as a rational function of the form $\frac{P_1(x,y)}{Q_2(x,y)}$. This proves the lemma.

Lemma II: *The five parameters of rigid motion are be uniquely determined by the parameters* p_i.

Algorithm IV: Equation (4.10) is the basis of a hough transform scheme to recover the motion parameters. The advantage of this scheme is that the constraint equation is linear in the synthetic parameters p_i. Once these

parameters are computed the five rigid motion parameters are uniquely determined. However this algorithm has the draw back that it requires an eight dimensional solution space. The next algorithm seeks to remedy this problem. It is based on the assumption that the optical flow field is available in the form of locally analytic functions. This enables us to obtain the first order spatial derivatives of the flow field, which are used to derive motion constraint equations.

Algorithm V: Differentiating equation (4.10) with respect to the retinal space coordinates we have two independent equations:

$$(yu_x - v - xv_x) + p_x v_y - p_2 u_t - p_4 + 2p_6 z - p_8 y = 0 \quad (4.14)$$

$$(u + yu_x - xv_x) + p_1 v_y - p_2 u_y - p_5 + 2p_7 y - p_8 x = 0 \quad (4.15)$$

The parameters in equations (4.14) and (4.15) are linearly independent when the depth function is not of the form given in lemma I. Selecting five suitable points we obtain two alternative sets of simultaneous equations in five unknowns. These can then be solved for the five motion parameters. Note, however, that when $p_x = \alpha_0 = 0$ then then equation (4.14) alone cannot be used for the computation. This is because the parameters $(P_1, P_2, P_4, P_\alpha, P_8)$ cannot then be used to solve for the five motion parameters. A similar restriction holds for equation (4.15) when $p_2 = y_0 = 0$.

Algorithm VI: *Motion parameters from structure and optical flow.*

When the structure of the moving surface is known, its motion is unambiguous. This method also reduces the dimensionality of the parameter space by isolating the rotational parameters. Two alternative constraint equations can be used here. In the first form spatial derivatives of the optical flow function are needed. This implies local analytic reconstruction of the flow function. In the alternative form of the constraint depth ratios are needed, implying reliable (and dense) measurement of surface normals.

From eq. (4.5) the expressions for the spatial derivatives of the optical flow (u,v) are obtained as:

$$\text{" . - - -f - (*o- «)-|i-|f-- «y + 2?x} \qquad (4.16.1)$$

$$\text{s--(».-»>"7r!r-«-7} \qquad (4.16.2)$$

$$\text{f, = - (*>- »);|rSf + /»» + ?} \qquad (4.16.3)$$

$$v_y = -\frac{W}{Z} \qquad y)\frac{W}{Z^2}\frac{\partial Z}{\partial y} - 2\alpha y + \beta x \qquad (4.16.4)$$

Substituting $(z_0 - x)\frac{}{Z}$ and $(y_0 - v)\sim ir\ ^{*n}$ the above equations from equation (4.5) we get:

$$u_s - v_y = (-u - \alpha xy + \beta(x^2 + 1) - \gamma y)\psi$$

$$+ (v + \alpha(y^2 + 1) - \beta xy - \gamma x)\rho + \alpha y + \beta x \qquad (4.17.1)$$

$$u_f = (-U - o\text{tx}y + 0\{x^* + l)\ -_7\ »')p - ax - 1 \qquad (4.17.2)$$

$$v_s = (-v - a\cdot(V^2 + 1)' + fixy + ix\}\psi + \beta y + 7 \qquad (4.17.3)$$

where $V = \frac{\frac{dZ}{dx}}{Z}$ and $\hat{P} = \frac{\frac{\partial Z}{dy}}{Z}$

Thus at every image location (x,y), a set of three linear independent equations involving the rotation parameters can be obtained. The functions $i>(x,y)$ and $p\{x,y)$ are computable from the surface orientation values $\frac{\partial Z}{\partial X}|_{s>r}$ and $\frac{\partial Z}{\partial Y}|_{s>r}$ (see Appendix B).

When it is not possible to measure derivatives of the optical flow, but the ratio of depths at any two image locations can be estimated, an alternative linear constraint equation can be derived involving only the rotation parameters. Consider two image points $(x_u y_x)$ and $(«_2, y_2)$ with depths $z\backslash$ and $*2$ respectively. The optical flow values at these points are $(«i, «i)$ and $(u_2, t/_2)$. The motion parameters are $\{U,V,W,a,0,i)$. Using equation (4.5) we have the following equations

$$u_x z_x - u_2 z_2 = (x_2 - x_x)\, W + z_x(- <*x_x y_x + /3(x_x^2 + 1) - iy_x) - z_2(- ax_2 y_2 + \beta(x_2^2 + 1) -$$

$$v_x z_x - v_2 z_2 = \{y_2 - y_x) + z_x(- a\{y_x^2 + 1) + 0 * \wedge + 7«i) - *a(\sim a(y_2^2 + 1) + 0x_2 y_2 + \gamma x_2)$$

Eliminating W from the above equations we have

$$l_{X2} a + m_{12} 0 + r i_2 7 + <_{i2} = 0 \qquad (4.18)$$

where

$$l_{X2} = x_x y_x y_2 - x_2 y_x * + x_x - x_2 + \frac{z_2}{z_1}(x_2 y_1 y_2 - x_1 y_2^2 - x_1 + x_2)$$

$$m_{12} = x_x x_2 y_x - x j^2 y_2 + y i - y_2 + \frac{z_2}{z_1}(x_x x_2 y_2 - a?_2^2 y_1 - y_x + y_2)$$

$$r_{12} = Z!X_2 + y_x y r - *\backslash^2 - V\backslash^2 + \frac{z_*}{\wedge}(- x_2^2 - y_2^2 + x_1 x_2 + y_1 y_2)$$

$$s_{12} = u_1(y_2 - y_1) - v_1(x_2 - x_1) + \frac{z_2}{z_1}(-u_2(y_2 - y_1) + v_2(x_2 - x_1))$$

If the surface normal value are available everywhere in a region enclosing two image points, then the depth ratio, $\frac{z_2}{z_1}$, (corresponding to those locations) can be estimated (of course, mathematically, it is possible to compute this ratio if the surface normal values are known along a path from the one image location to the other). Consequently, each pair of image points gives rise to a linear constraint in the rotation parameters. Thus by a suitable choice of three pairs of image points we can uniquely solve for the rotation parameters and subsequently the translation parameters $(\frac{U}{W}, \frac{V}{W})$ (see Appendix A).

The novel feature of the above algorithm is that it can combine shape and motion information under two different conditions:

(1) In the first case the optical flow field has been measured sufficiently 'densely' to enable local reconstruction of the flow field. This enables the first order spatial derivatives of the flow field to be estimated. Then at all retinal points where the surface normals are known, we can locally solve for the rotation parameters by means of a set of three linear constraint equations.

(2) Alternatively, if the flow measurements are not dense, but the shape measurements allow reconstruction of the depth function (up to a constant scale factor), then again locally we obtain linear constraints in

the rotation parameters (e.g. equation (4.18)).

This means that in any image neighborhood, full reconstruction of either shape or image motion, helps to recover both structure and motion. The schematic diagram of the algorithm is given in figure 4.5.

The implemented algorithm uses the constraint equation obtaining $ and p from equations (4.17), to obtain a cubic polynomial equation in the three rotation parameters. The optical flow and its first spatial derivatives are measured and the cubic constraint is used to estimate the rotation parameters by the hough transform technique. So, although the nonlinearity remains, the dimension of the parameter space is reduced, which reduces the size of the search space. The effect of parameter space quantization for algorithm VI can be seen in table 4.2.

4.5. Conclusions

This chapter reported the results obtained experimentally using motion interpretation algorithms based on the constraints developed in chapter three. The hough transform was chosen as the preferred scheme for implementing the algorithms since it is implementable by simple massively parallel architectures [31] • In the case of linear constraints least square error

Quantization	Computed Axis			Error
N	x	y	z	<%)
5	.25	.47	.84	7
3	.17	.81	.56	36
2	.00	.67	.75	30
1	.00	.36	.93	36

Table 4.2 Error in determination of axis of rotation

minimization methods can be applied, however these techniques are no longer appealing when the constraint equations are nonlinear. The hough Transform technique extends to the nonlinear case.

This chapter also introduced the notion of a hierarchic hough transform scheme where a coarse to fine refine strategy was seen to work well with the nonlinear constraint equations that arise in relation to rigid body motion.

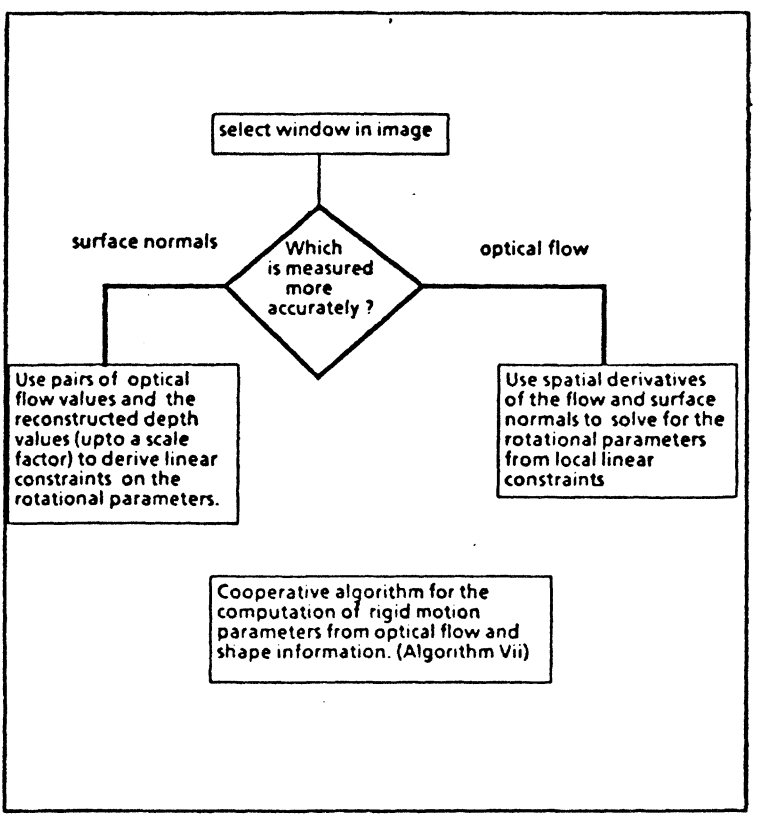
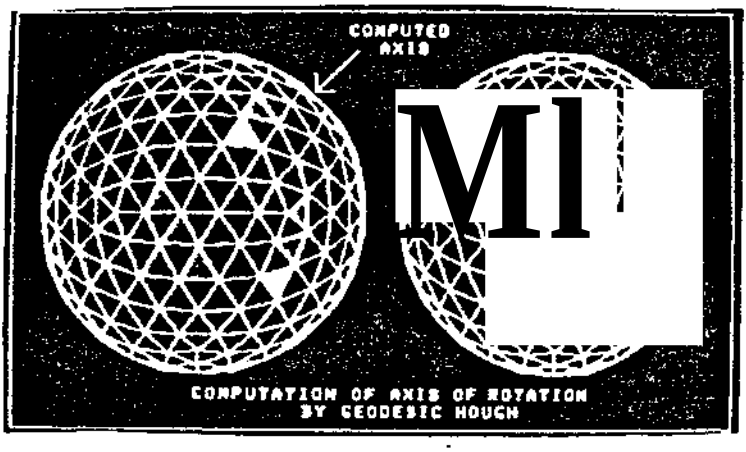
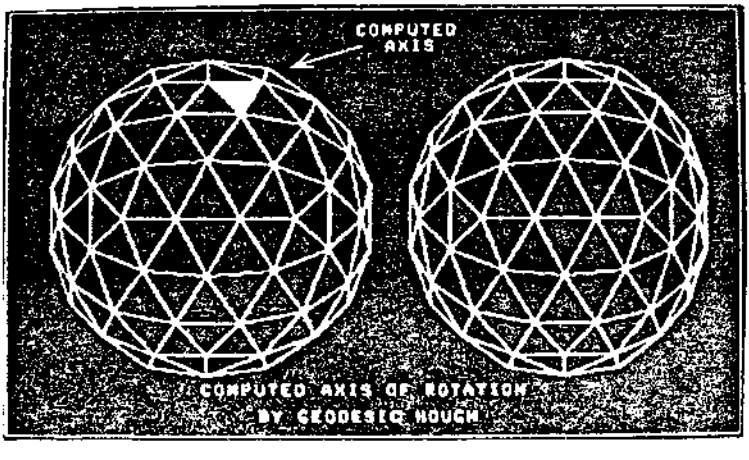


Figure 4.5 An Adaptive algorithm for determining rotation

PLATE 4.1. Hough Transform in rotational subspace.

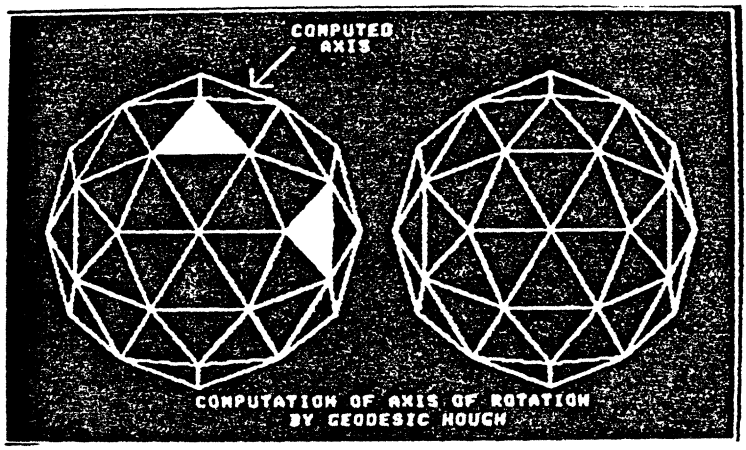


A. Resolution parameter N = 5

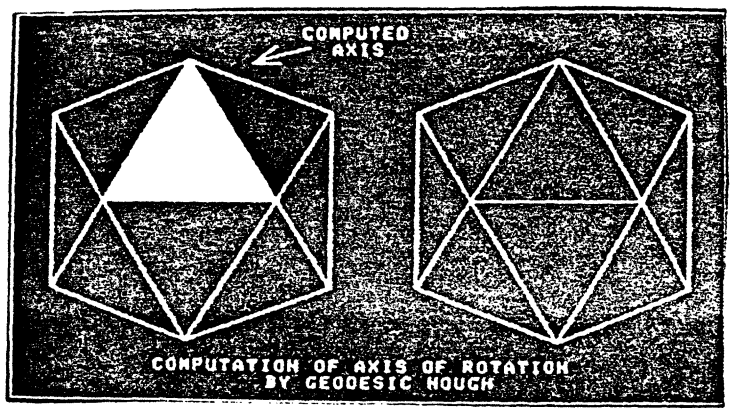


B. Resolution parameter N = 3

C. Resolution parameter N = 2



D. Resolution parameter N = 1



Chapter Five

Active Navigation

Egomotion Perception by the Tracking Observer

5.1. Introduction

The perception of rigid motion finds application, in many areas. Some of these have been mentioned previously. One of the more important ones is the computation of egomotion parameters with the help of visual stimuli. These parameters help in registering the observer's motion with respect to the environment and are prerequisites to navigation. The problem that is addressed in this chapter, is termed the *Visual Navigation* problem. The goal here is to devise means of computing the *Egomotion* parameters of a moving observer, from visual data.

The *Passive Navigation* approach [19] deals with egomotion parameter computation, when the moving observer carries an optical imaging device(s) which obtains time-varying imagery of the surrounding scene. The computation usually assumes that image motion (e.g. optical flow) has been

computed previously, and is available as input to the perceptual process. Recently, there have been attempts to do "direct computation" [64] or "motion without correspondence" [3] based on restricted situations like planar moving surfaces or purely translational motion. These approaches are novel, and certainly merit attention, but are of a preliminary nature and need further study. Egomotion perception under the monocular passive technique is handicapped by nonlinear constraint equations. In addition, the difficulty of the computational task is compounded by the fairly large number of unknown parameters to be determined. Therefore, since the equations cannot be decoupled or simplified, iterative search techniques or parameter space histograming (hough transform) have to be used in the parameter determination. In this chapter, an alternative approach to computing Egomotion is proposed. This technique requires the moving observer to visually track an environmental feature. Our term for this proposed method of egomotion perception is *Active Navigation*. It must be clarified, however, that the sensing method used is not active (e.g. laser or ultrasound ranging), but the perceptual system operates in a closed loop fashion with active feedback from the image motion computation module. The various advantages of the method are discussed. An analysis of the geometry of this particular situation is examined to outline how closed form solutions for the parameters may be obtained.

The strategy advocated in this chapter calls for visual tracking and combination of information from stereo image pairs. This approach is adopted based on a number of experimental simulation studies reported in the previous chapter and also in [28, 84], which indicate the difficulty of the passive monocular approach. The stereo motion approach is also under investigation elsewhere [50, 91], and the possible employment of active tracking to facilitate navigation has been suggested by visual psychologists [22].

It will be shown that when the observer is able to track a prominent feature point in the imaged scene, the task of navigation is facilitated since it is easier to compute egomotion parameters, compared to the non-tracking case* The emphasis in this chapter is on the mathematics governing the imaging equations that are obtained while the system is tracking. To track, the system must have some way of measuring the error in the retinal signal. Ways of doing this are discussed in section 5.2.

The outline of this chapter is as follows:

1 Error velocity measurement to correct tracking drift is discussed in light of the primate pursuit system.

2. A general form of the relation between 3D velocity parameters and retinal optical flow is derived. In previous derivations of this relation [52] the origins of the body centered coordinate frame and viewer centered coordinate frame are taken to coincide at the instant of

measurement. Using the general representation it is shown why a monocular observer, who is able to track an environmental feature point, has to contend with a smaller number of velocity parameters.

3. The analysis extends naturally to stereo imaging situations, where it is shown that, by combining measurements from both eyes, a linear equation in two unknowns is obtained.

4. The above constraint is applied with all possible stereo correspondences in a small neighborhood, so as to minimize the square error. This least square error technique is seen to work well on simulated data, even with the addition of 10-20 percent noise.

5. A new set of constraint equations are derived for the tracking observer, which allow closed form solution of the egomotion parameters. Simulation results are described and implementational issues for integrating this module in the overall motion interpretation scheme are discussed.

5.2. Target selection via Velocity Channels

The key assumption is that the alignment of the camera axes are controllable by the system itself. In this case, as the system moves in the world, the orientation of the camera is continually adjusted. This adjustment is dependent upon the two dimensional motion perceived on the retina.

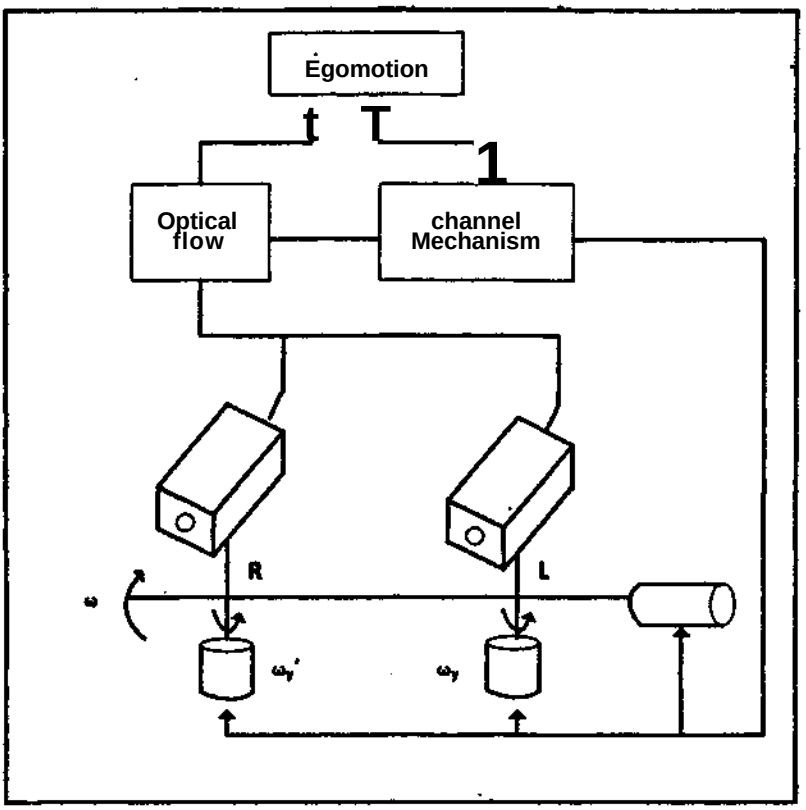


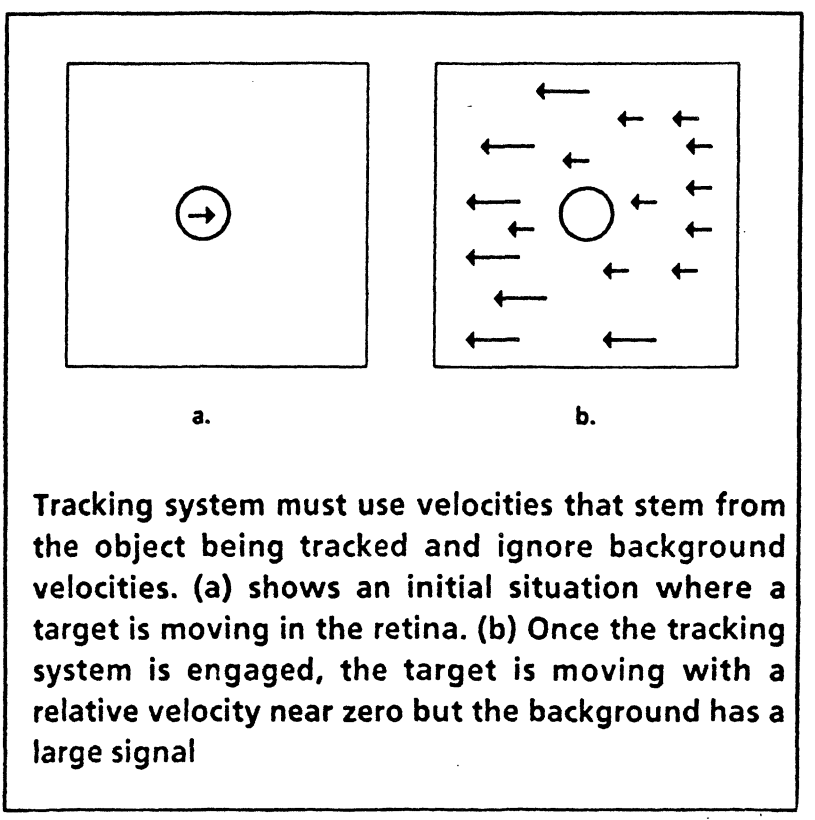
Figure 5.1 Hie Tracking Mechanism

In the tracking system the problem can be seen as, given the image of a target environmental point, to generate control signals that will *foveate* the target The block diagram of a system for accomplishing this can be schematized as shown in figure 5.1. The first and most important point to make **is that the system can be adequately modeled by servomechanism concepts. It is relatively easy to see how to generate the kinds of motor commands for the two movement systems to produce the observed** behavior. This of course assumes that the target point is identified.

Target identification is a central issue: in a complicated motion field, how can the target velocities be easily identified ? This **is** a basic

subproblem in tracking using velocity sensing and is captured by figure 5.2.

Our answer to this question uses the notion of global flow field vectors. Such vectors respond to velocities in every part of the optical flow field. In other words, if we visualize the optic flow field as a four dimensional parameter space (x,y,u(x,y),v(x,y)), The global flow field sums all the different flow vectors in a two dimensional (u,v) parameter space. Detectors form a distinct set whose sensitivities are organized into channels. In the case of a particular flow field, some channels will typically respond to it and others will not. Figure 5.3 shows how the channel concept can be



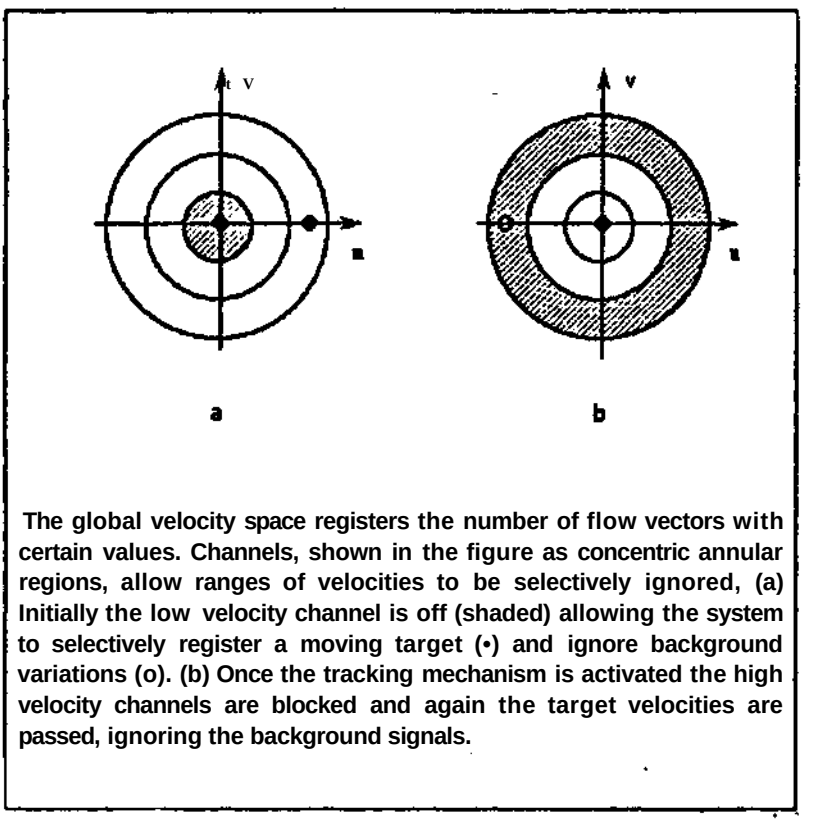
Tracking system must use velocities that stem from the object being tracked and ignore background velocities. (a) shows an initial situation where a target is moving in the retina. (b) Once the tracking system is engaged, the target is moving with a relative velocity near zero but the background has a large signal

Figure 5.2 Target Identification

utilized.

We claim that with this abstract flow channel model the problem becomes one of determining which of the channels should be used for the eye movement control system. This means that a mechanism is needed to switch the appropriate channels into the servo system. Note that this technique uses a spatially distributed detector array. Our contention is that it is appropriate to average the flow field over this subset.

A mechanism to switch the detectors on once the appropriate ones have been identified is simple to understand, so we will concentrate on



The global velocity space registers the number of flow vectors with certain values. Channels, shown in the figure as concentric annular regions, allow ranges of velocities to be selectively ignored, (a) Initially the low velocity channel is off (shaded) allowing the system to selectively register a moving target (•) and ignore background variations (o). (b) Once the tracking mechanism is activated the high velocity channels are blocked and again the target velocities are passed, ignoring the background signals.

Figure 5.3 Concept of the Velocity Channel

identifying the ideas behind selecting the right detectors. The general way that this is done is by a feed-forward mechanism that determines some selection criterion. The different kinds of criteria are important, so it is useful to categorize them.

1. *Extrinsic features.* This method uses some other feature, say color, that also has spatially organized detectors. To track a red object, the detectors that register red are used to select the spatial component of the velocity detectors. All such detectors with the appropriate correspondence are used.

2. *Intrinsic Features.* This method uses some particular range of values for the flow field itself, say all values over a certain velocity magnitude. To track an object, all the detectors that satisfy the intrinsic criterion are switched into the movement control system.

These distinctions are important as they correspond to two different types of tracking situations. In navigation, where the entire spatial field is moving, an extrinsic feature is appropriate. In pursuing a small target, that target is usually moving differently with respect to the background, so an intrinsic feature may be appropriate.

5.3. Measuring Egomotion

5.3.1. Background

Consider first, the monocular imaging situation where a sensor is moving relative to a static scene. The co-ordinate frame (X,Y,Z) is fixed to the sensor (see figure 5.4). The viewing direction is along the positive Z-axis.

The analysis presented here assumes a rotating and translating observer moving in a static environment. However, since the velocity parameters that characterize the motion are all relative to the observer's frame of reference, the analysis per se, is not affected by multiple moving objects. The analysis assumes the velocity representation for the motion parameters.

The reference coordinate frame is fixed to the observer. There is another coordinate frame fixed at the point 'S' on the body (see figure 5.4). The point S has the velocity $\mathbf{T}_s = (U_s, V_s, W_s)^T$. At the time of observation the reference and the body frame axes are parallel to each other. The rotational velocity of the body is given by the vector $\Omega = (\alpha, \beta, \gamma)$. The 3D velocity of a point $P = (X, Y, Z)$ on the body is given by the equation

$$\dot{\mathbf{X}} = \mathbf{T} + [R]\,(\mathbf{X} - \mathbf{X}_s) \tag{5.1}$$

where $\mathbf{X}_s = (X_s, Y_s, Z_s)$ denote the position of the body origin 'S' , and $\dot{\mathbf{X}}$ denotes the 3D velocity of P (the 'dot' operator is used throughout to signify differentiation with respect to time), also

$$[R] = \begin{bmatrix} 0 & -\gamma & \beta \\ \gamma & 0 & -\alpha \\ -\beta & \alpha & 0 \end{bmatrix}$$

The image formation is modeled by the perspective projection model (see chapter three). The projection of a point $P=(X,Y,Z)$ is denoted by $p=(x,y)$. The projective relation is

$$(x,y) = (\frac{fX}{Z}, \frac{fY}{Z}) \tag{5.2}$$

The constant f is the focal length of the imaging system. It is the distance separating the nodal point of the camera (or eye) and the image plane, moving along the optical axis (i.e. Z axis). In subsequent steps the constant f is assumed to be unity. The velocity of image points in the 2d image space is called optical flow. The relations between the 2D and 3D velocities are obtained by differentiating the equation (5.2) and substituting from equation (5.1).

$$\begin{aligned} u = \dot{x} &= \frac{U_t - xW_t}{Z} - \alpha[xy - x\frac{Y_t}{Z}] + \beta[1 - \frac{Z_t}{Z} + x^2 - x\frac{X_t}{Z}] - \gamma[y - \frac{Y_t}{Z}] \\ v = \dot{y} &= \frac{V_t - yW_t}{Z} - \alpha[1 - \frac{Z_t}{Z} + y^2 - y\frac{Y_t}{Z}] + \beta[xy - y\frac{X_t}{Z}] + \gamma[x - \frac{X_t}{Z}] \end{aligned} \tag{5.3}$$

When the origin of the body coordinate frame coincides with the reference or observer coordinate frame then $X_t = Y_t = Z_t = 0$, and $T = T_0 = (U,V,W)$, which simplifies the equation for optical flow to give :

$$u = \frac{U - xW}{Z} - \alpha xy + \beta(x^2 + 1) - \gamma y \tag{5.4.1}$$

$$v = \frac{V - yW}{Z} - \alpha(1 + y^2) + \beta xy + \gamma \tag{5.4.2}$$

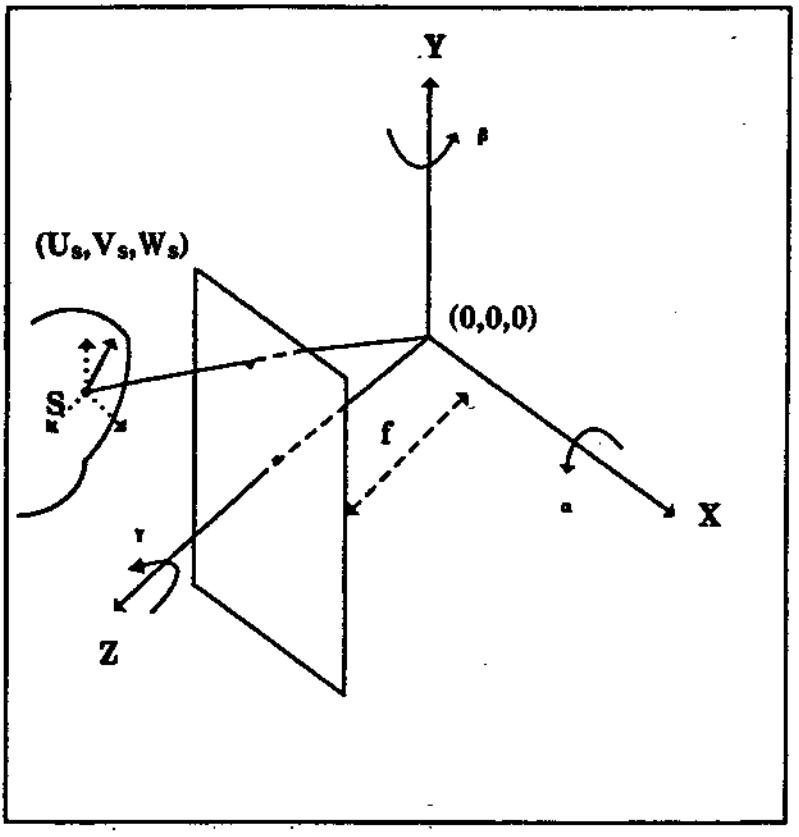


Figure 5.4 Imaging Geometry and motion representation

The above pair of equations embodies the constraint that the optical flow (u,v) imposes upon the the parameters of rigid motion. Thus all an observer has to do to determine where he is going is to measure the retinal velocity pattern and then use the above pair of equations applied at least five points [13, 67, 84], to determine the 3D velocity of egomotion. Note that there are six velocity, components (Le. three for translation and three for rotation). Unfortunately however, all the six parameters cannot be computed by monocular visual data. This is because of the depth term 'Z' that occurs in the above pair of equations. The depth introduces a scaling effect, whereby other things being equal, multiplying the translational

components and the depth by the same constant factor leaves the perceived retinal motion unchanged. Thus for example an object at a certain distance, translating with a certain speed generates the same optical flow field when it is twice as far away and traveling in the same direction with twice the speed.

Thus the monocular observer, lacking depth information, must eliminate the depth factor from the optical flow constraints. This will then imply that the observer's translation can only be determined up to a scale factor. Thus the number of egomotion parameters of interest are five - pertaining to the direction of translation and the rotation.

When the depth variable is eliminated from the above equations we have

$$\frac{g_0 - \wedge}{Vo - V} \quad \begin{matrix} u + axy - 0\{x^2 + .\ 1) + 7? \\ v + a(y^2 + 1) - fixy - 7^* \end{matrix} \qquad \begin{matrix} ,g\,5\,V. \\ \backslash\ '\ ) \end{matrix}$$

where $(x_{Oi}y_Q) = (\frac{U}{W}, \frac{V}{W})$ represents the direction of translation of the observer's coordinate frame.

The above constraint equation demonstrates the difficulty of motion computation for a monocular observer. It is nonlinear as well of high dimensionality, both this properties in conjunction make the problem difficult ([13, 28, 52, 53, 70, 84]).

5.3.2. The tracking Advantage

It will now be shown that in case the monocular observer can discern a distinguishing feature or mark on the observed surface then the perception

problem becomes simpler* Suppose that the surface in view has an easily distinguishable and localized feature at point 'S' whose corresponding image location is (x_9, y_9). In this case we can shift the body origin to the point S and rewrite the optical flow equations as in (5.3). In addition only translation

$$u(x_s, y_s) = \frac{r.-Lw.}{Z_s} \qquad (5.6)$$

Combining equations (5.6) and (5*3) one obtains:

$$u = \frac{u_s + (y - y_s)W'}{Z'} + \frac{axy - P(1+x^2)}{Z'} + \frac{1Vt}{axy} + P(1+X^2) \sim 7y \quad I^5\text{-}7\text{-}1)$$

$$v = \frac{v_s + (y - y_s)W'}{Z'} + \frac{\alpha(yy_s+1) - \beta x\, y - \gamma x_s}{Z'} - \text{or}(1\text{-}fy^2) +)9xy + 7X \quad (5.7.2)$$

where the 'prime[1] operator signifies scaling by Z_{99} i.e. $W_9* = \frac{W_9}{Z_s}$. Note that the translational parameters with respect to the observer's frame (i.e. the observers actual translation) are related to the body centered translational parameters by

$$\begin{aligned} U' &= U_s' - \beta + \gamma y_s \\ V &= V + a - 7*. \\ W' &= W_9' \sim a y_9 + fix. \end{aligned} \qquad (5.8)$$

The above analysis illustrates the fact that given the ability to estimate the projected velocity of a localized feature accurately, the constraint equations reduce in dimensionality by one.

A similar result may be obtained, as can be expected, when the moving observer is able to track a single feature point so that it appears stationary

on the retina at position (0,0). In this case we assume that the tracking motion consists of rotations about the axes that are orthogonal to the line of sight or the optical axis of the lens. The tracking motion is a rotation $(\omega_x,\omega_y,0)$, which is superimposed upon the actual parameters of motion.

Let $S=(0,0,Z_0)$ be the spatial coordinates of the point being tracked. Assume that the observer can track an environmental point and hold it steady on the optical axis (Z axis). Therefore the optical flow field will have a singularity at the origin of the retinal frame, where the flow value is zero. At the time of observation, the tracked point tends to move along the observer's optical axis (figure 5.5).

Consider an observer moving with translation (U,V,W) and rotation (α,β,γ). Then, if the body frame origin is taken to be at S, from equation (5.8), remembering that $U_t=V_t=0$:

$$\begin{aligned} U' &= \frac{U}{Z_0} = -B \\ V' &= \frac{V}{Z_0} = A \\ W_t &= W \end{aligned} \tag{5.9}$$

furthermore the optical flow equation (5.3) becomes:

$$\begin{aligned} u &= \frac{-xW}{Z} - Axy + B\left[1 - \frac{Z_0}{Z} + x^2\right] - \gamma y \\ v &= \frac{-yW}{Z} - A\left[1 - \frac{Z_0}{Z} + y^2\right] + Bxy + \gamma x \end{aligned} \tag{5.10}$$

where $A=\alpha+\omega_x$ and $B=\beta+\omega_y$.

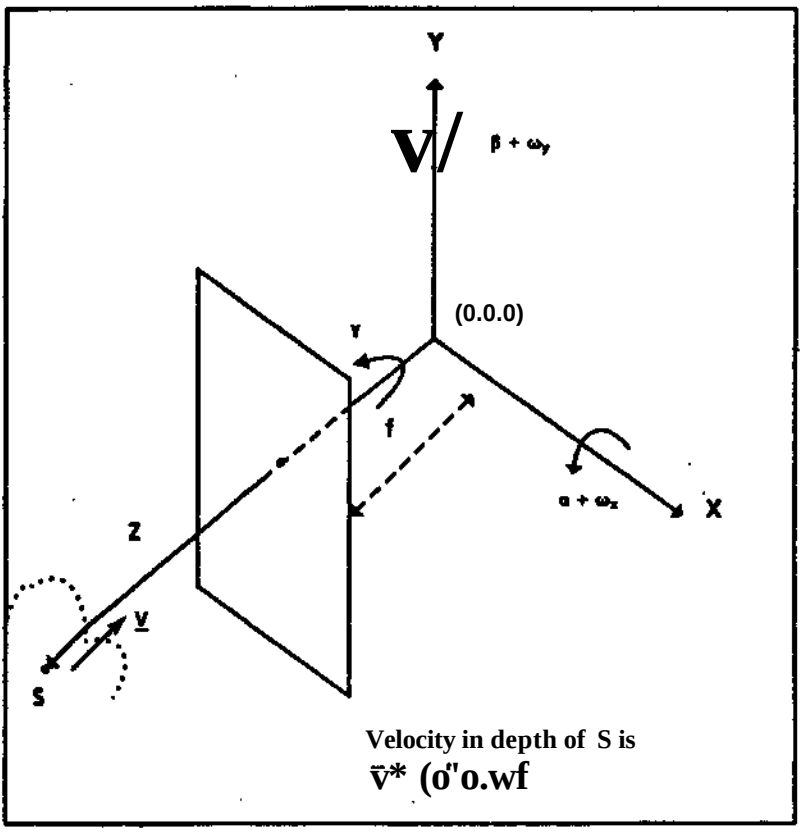


figure 6.5 Monocular Tracking

Eliminating Z from the above we have:

$$\frac{u + Axy - B(x^2 + 1) + \gamma y}{v + A\{1 + y^2\} - Bxy - {}_7x} = - \frac{B + xW}{A - yW'} \qquad \text{(5-H)}$$

where $1^{\wedge}{}' = -^{\wedge}_{0}$.

The constraint equations derived above are similar in form to equation (5.5). However, in this case the dimensionality of the parameter space has **been reduced from five to four** ***without increase in the degree of nordinearity of the constraint*** It is important to note that the observer can determine his direction of translation since from equation (5.9) we have

$$U' = \frac{U}{Z_0} = -B$$

$$V' = \frac{V}{Z_0} = A$$

Thus even without explicitly measuring his tracking motion, the observer can determine the scaled translation (U', V', W'). We next examine the constraint equation (5.10) and show how it may be used to actively compute the direction of translation.

5.4. Stereo tracking

It can be expected that stereoscoping viewing can simplify the task of motion perception. Binocular imaging system does introduce a new complication in that in addition to the task of retinal motion estimation, one must also accomplish stereo fusion. However stereo fusion is a simpler task than optical flow estimation, and a recently published algorithm is reportedly able to handle this task reasonably satisfactorily [69]. The advantages of stereo imaging for analyzing motion are:

1. The motion constraint equation is linearized.
2. Tracking an environmental feature point greatly simplifies motion computation under stereo imaging conditions.
3. The epipolar constraint is a powerful aid in handling the "correspondence problem" for both stereo fusion as well as retinal motion estimation. (In this section the reason for this will be sketched, but it will not be detailed)

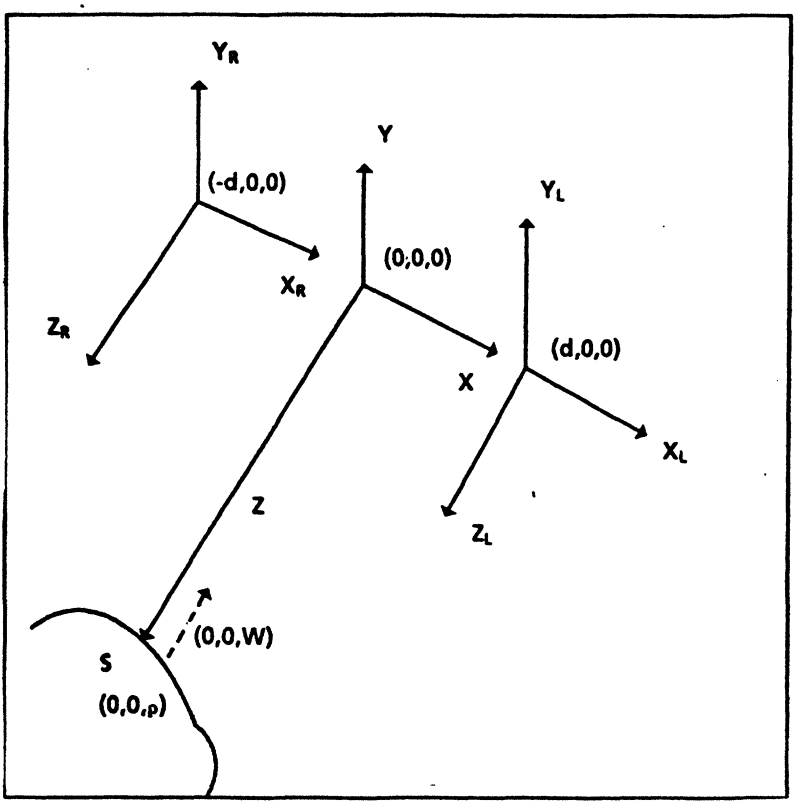


Figure 5.6 Tracking in stereo

5.4.1. Tracking in Stereo with Parallel Camera Axes

The monocular imaging geometry described previously is augmented by two other coordinate frames located at the points $(d,0,0)$ (the left eye frame) and $(-d,0,0)$ (the right eye frame) respectively. The central frame can be imagined as the "head frame" and the two other frames as the camera or "eye" frames. The situation is depicted in (figure 5.6). In this scheme there is no vergence between the two eye frames (rather the eyes verge at infinity). This means that the corresponding axes of all the coordinate frames are parallel. Furthermore, it is assumed that the frames are rigidly attached to each other.

The tracking action is with respect to the head frame. Now if the head frame is tracking a feature point S_h at (0,0,p) then its image on the left and right eye frames are at (-e,0,0) and (e,0,0) respectively. The relation between the depth p and t is

$$p \quad \frac{2\pm}{2e} S = \pm \quad p$$

Once again for simplicity of explanation, consider the relative motion between the observer's head frame and the observed rigid scene, as due to egomotion. The motion parameters are the translational velocity *(U,V,W)* and the rotational velocity (<*,£,7). The observer's tracking movement is confined, as before, to the rotation (w_s,o>$_f$,0), with respect to the head frame. The tracking motions executed by the the eye frames include this rotation *plus* translations in depth of - dw_9 and dw, respectively.

Consider an image location *{$x_{l9}y$)* in the left frame, and its stereo pair (x_r,y) in the right frame. The disparity is given by

$$\frac{}{Z}$$

where Z is the depth **of** the point in space giving rise to the stereo pair.

The motion parameters **are** as **before** *(U,V}W)* **and** (a,£,7), with respect **to a hypothetical head frame located between the two stereo coordinate** frames. The head frame is assumed to track the environmental feature S_K (the subscript refers to the fact that the nomenclature is with respect to the head frame). Therefore equations (2.9) hold. The motion parameters with

respect to the stereo frames are:

$$L : \mathrm{Ti} = \mathrm{T!}^{b} + \mathrm{T},^{*} \qquad \mathrm{n}_{(} = \mathrm{n},^{b} + \mathrm{ni}^{*}$$
$$R : \mathrm{T_F} = \mathrm{T_r}^{b} + \mathrm{T}^{*}_{r} \qquad \mathrm{Q}_{,} = \mathrm{Q}_{r}^{b} + \mathrm{n}_{r}^{*}$$

where the subscripts / or r refer to the left or right frames, and the superscripts *b* or *tr* refer to body parameters (or representing actual motion) and motion induced due to the tracking motion respectively. These components can be expanded to

$$\mathrm{T},^{b} = (\ U,\ V + td,\ W- \quad fid) \qquad \mathrm{fli}^{b} = (\ \mathrm{«}\ .\ *\ .\ 7)$$
$$\mathrm{T}_{r}^{b} = (\ U,\ V-\ {}_{1}d_{9}\ W + pd) \qquad \mathrm{n}_{r}^{b} = (\ \mathrm{a}\ ,\ \wedge\ ,\ {}_{7}\ )$$

and

$$\mathrm{Tf}-(\mathrm{O_fO_f}-\ \mathrm{«_f d}) \qquad \mathrm{fli}^{fc}=(\mathrm{«_{lf}«i},0)$$
$$T?.=*\{0,0_{t}W_{9}d) \qquad n?=(*,\ ,\mathrm{«}_{f\ f}0)$$

It can be seen that the motion of the tracked point, is given as $T_9 = (\ 0\ ,\ 0\ ,\ W\ )$ in both the left and right frames. The rotation of these frames is also the same, namely (A , B , 7). Finally, the tracked point is located with respect to the two frames as: $5/ = (\ -d\ ,\ 0,\ p\ )$ and $S_r = (\ d\ ,\ 0\ ,\ p\ )$. Therefore from equation (2.3) we get the optical flow constraints for the left eye as:

$$v_l = \frac{-x_l W}{Z} - Axy + B\left(1 - \frac{\rho}{y^{+}+*/^{2}} + -\frac{x_l d}{j-}\right) - IV$$
$$v_l = -\frac{yW}{Z} - A\left(1 - \frac{\rho}{Z} + y^{2}\right) + B\left(x_l y + y\frac{d}{Z} + {}_{1}\left(z_{t} + \frac{d}{Z}\right)\right.$$

where $\mathrm{A = a + w}$, and $\mathrm{B = j5 + w_r}$ In the above equation, making the substitution ($Z = \_\_\_\_\_$) we have:

$$u_l = x_l\phi + B(1 - \frac{\rho}{Z}) - \gamma y$$
$$v_l = y\phi - A(1 - \frac{\rho}{Z}) + \gamma\frac{(x_l + x_r)}{2} \tag{5.12}$$

where $\phi = -\frac{W}{Z} - Ay + B\frac{x_l + x_r}{2}$. similarly the optical flow for the right eye is given by:

$$u_r = x_r\phi + B(1 - \frac{\rho}{Z}) - \gamma y$$
$$v_r = y\phi - A(1 - \frac{\rho}{Z}) + \gamma\frac{(x_l + x_r)}{2} \tag{5.13}$$

From the above equation we get:

$$\phi = \frac{u_r - u_l}{x_r - x_l}$$

which leads to a constraint equation in two parameters:

$$v_r = v_l = y\frac{u_r - u_l}{x_r - x_l} - A(1 - \frac{\rho}{Z}) + \gamma\frac{x_l + x_r}{2} \tag{5.14}$$

This with stereo tracking it is possible to obtain a linear constraint in two unknowns at every point of measurement.

5.4.2. Tracking with Convergent Stereo Imaging

In this case the optical axes of the two cameras converge onto a point in the environment that is being tracked. The geometry is illustrated in (figure 5.7). We will generally deal with the left coordinate frame, with respect to which the various quantities will be written as in the monocular case. When we need to reference the quantities with respect to the right frame these will be written primed (e.g. x'). The tracking motion involves three independent rotational velocities (ω,ω_y,ω_y'). The rotation ω is about

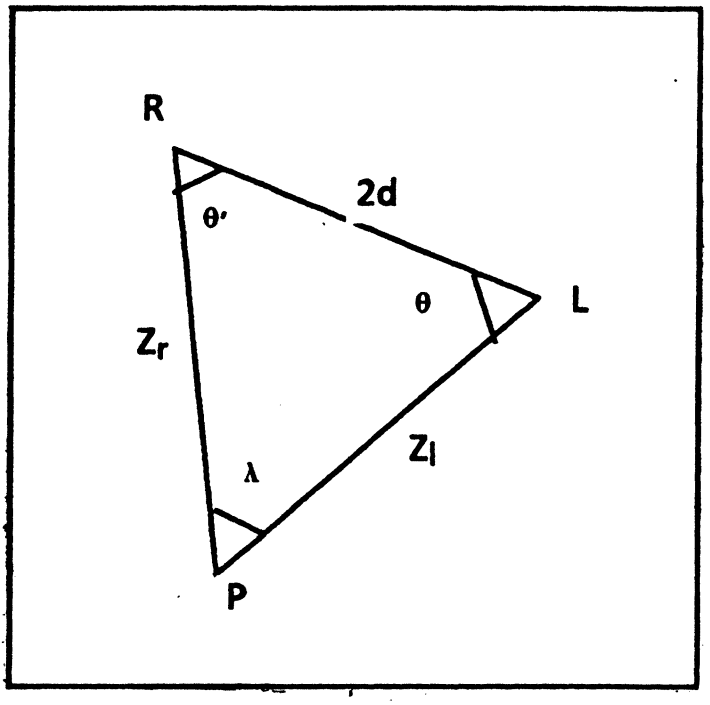


Figure 5.7 Binocular Convergent Tracking

the baseline RL of the imaging system (figure 5.7). Hence the tracking motion of the left frame is given by ($\omega_x = -\omega\sin\theta$, ω_y, $\omega_z = \omega\cos\theta$). If Z_0 is the depth of the tracked point in the left frame then:

$$\frac{2d}{\sin(\theta+\theta')} = \frac{Z_0}{\sin(\theta')} = \frac{Z_0'}{\sin(\theta)}$$

Thus we can write:

$$Z_0(t) = 2d\frac{\sin(\theta')}{\sin(\theta+\theta')} = F(\theta(t),\theta'(t))$$

also

$$\dot{F} = 2d\left[\frac{\dot{\theta}'\cos\theta'\sin(\theta+\theta') - (\dot{\theta}+\dot{\theta}')\sin\theta'\cos(\theta+\theta')}{\sin^2(\theta+\theta')}\right]$$

which simplifies to

$$\dot{F} = 2d\left[\frac{\dot{\theta}'\sin\theta - \dot{\theta}\sin\theta'\cos(\theta+\theta')}{\sin^2(\theta+\theta')}\right] \tag{5.15}$$

Differentiating the above relation with respect to time once more, we have:

$$\ddot{F} = 2d\left[\frac{\ddot{\theta}'\cos\theta'}{\sin(\theta+\theta')} - \frac{(\ddot{\theta}+\ddot{\theta}')\sin\theta'\cos(\theta+\theta')}{\sin^2(\theta+\theta')}\right]$$

$$+ 2d\left[\frac{(\dot{\theta}+\dot{\theta}')^2\sin\theta' - (\dot{\theta}')^2\sin\theta'}{\sin(\theta+\theta')}\right]$$

$$- 4d\left[\frac{(\dot{\theta}+\dot{\theta}')\dot{\theta}'\cos\theta'\cos(\theta+\theta')}{\sin^2(\theta+\theta')} - \frac{(\dot{\theta}+\dot{\theta}')^2\sin\theta'\cos(\theta+\theta')}{\sin^3(\theta+\theta')}\right]$$

simplifying the above leads to

$$\ddot{F} = 2d\left[\frac{\ddot{\theta}'\sin\theta - \ddot{\theta}\sin\theta'\cos(\theta+\theta')}{\sin^2(\theta+\theta')}\right]$$

$$+ 2d\left[\frac{\dot{\theta}(\dot{\theta}+2\dot{\theta}')\sin\theta'}{\sin(\theta+\theta')}\right]$$

$$- 2(\dot{\theta}+\dot{\theta}')\cot(\theta+\theta')\dot{F} \tag{5.16}$$

Let the motion of the observer be described by the translational velocity $T=(U,V,W)$ and rotational velocity $\Omega=(\alpha,\beta,\gamma)$. These parameters are defined with respect to the point L in the body, which also happens to be the origin of the left coordinate frame. The tracking motion of the system consists of three independent rotations with respect to the observer. These three rotations correspond to the three motors in figure 5.1. The angular velocity ω corresponds to the rotation of the plane PRL about the axis LR. The other two angular velocities are $\dot{\theta}$ and $\dot{\theta}'$, which affect the left and right coordinate frames respectively. Let the sense of ω be positive in the direction from L to R. Then the tracking angular motion of the left frame

with respect to the observer is given by a>t and that of the right frame is o>t'. Note that we will the express all the motion parameters measured in a frame with respect to basis vectors defined in that frame. Therefore:

$$\omega\sin\theta,\ \dot{\theta},\ \omega\cos\theta)$$
$$o/_t' = (\ \_\ a;\sin 0',\ \dot{\$}\backslash\ -\ a/\cos(')$$

If the rigid motion parameters with respect to the right coordinate frame are given by the translations! velocity T and the rotational velocity fit' then we have:

$$\mathbf{T}' = R_{\lambda}\cdot(\mathbf{T} + \Omega \times \rho)$$
$$\Omega' = R_{\lambda}\cdot\Omega$$

where V denotes the vector product and * denotes matrix multiplication. In addition, the rotation matrix R_x expresses the transformation due to the rotation by X =*• - (0 + f) between the left and right frames and is

$$R_{\lambda} = \begin{bmatrix} \cos X & 0 & -\sin\backslash \\ 0 & 1 & 0 \\ \sin X & 0 & \cos X \end{bmatrix}$$

Now from equation (5.9) we have:

$$\begin{aligned} &\dot{F}(t) = W \\ &U + (\beta + \omega_{\rho})F(t) = 0 \\ &V - (<^* + w,)F\{t) = 0 \end{aligned} \tag{5.17}$$

Observe that the above equations involve five unknown motion parameters. If we now differentiate these equations we have

$$\begin{aligned} &\ddot{F}\{t) = \dot{W} \\ &\dot{U} + \dot{p}F\{t) + \dot{f}i\dot{F}(t) + w\backslash F\{t) + w_f/(0 = 0 \\ &\dot{V} - \dot{a}F(t) - o\dot{t}\dot{F}\{t) - u!_9F\{t) - u>,\dot{F}\{t) = 0 \end{aligned} \tag{5.18}$$

Although here we consider a rectilinearly moving observer, the translational

velocity *{U,V,W)* undergoes change due to the rotation of the frame in which the observations are made* Thus we obtain:

$$\dot{V} = -(a + a_{;,})\, W + (7 + a/_s)\, U$$
$$\dot{W} = \{a + w_s)V - (/?+a;_f)tf$$

Similarly the rotational velocity (0,^,7), undergoes change due to the tracking motion, as follows:

$$\dot{a} = a;_r 7 - w_M p$$
$$\dot{f}t = -\, a;_s 7 + a;,a$$
$$\dot{7}=0$$

Introducing the parameters $A = a + w_t$, $B = 0 - w_9$ and $C = 7 + w_s$, substituting for $\dot{U}_9$ $\dot{V}$, $\dot{W}$, $\dot{a}$ and $\dot{p}$ from the above relations, and replacing U and V from (5.17), we have, from the last two equations in (5.18)

$$\frac{-2B\dot{F}(t) + Aw,F\{t) + u^{>}_{w}F\{t)}{\omega_s)F(t)}$$

and

$$c = \frac{2A\dot{F}(t) \sim Bu,F(t) + \dot{\omega}_s F(t)}{-(B + \omega_s)F(t)}$$

Finally eliminating C from the above pair of equations and using the remaining equation of (5,18) we obtain the pair of independent equations:

$$A\,t + s\,i \sim 7\left(\frac{\ddot{F}(t)}{0}\right)'' \sim ' 1 \qquad (5,19\#1)$$

$$\langle j\rangle_2 A + 4\rangle_Z B + \langle f\rangle_A = 0 \qquad (5.19.2)$$

where

$$\phi_2 = 2\omega_z \dot{F}(t)F(t) + \dot{\omega}_z F^2(t) + \omega_y \omega_z F^2(t)$$
$$\phi_3 = 2\omega_y \dot{F}(t)F(t) + \dot{\omega}_y F^2(t) - \omega_z \omega_z F^2(t)$$
$$\phi_4 = (\omega_z \dot{\omega}_z + \omega_y \dot{\omega}_y)F^2(t) + 2\dot{F}(t)\ddot{F}(t)$$

From equation (5.19) we obtain two sets of solutions for the motion parameters. Eliminating the parameter B we have:

$$aA^2 + bA + c = 0 \tag{5.20}$$

where $a = \phi_2^2 + \phi_3^2$, $b = 2\phi_2\phi_4$ and $c = \phi_4^2 - \phi_1\phi_3^2$

To summarize, the solution method consists of obtaining the solutions to the pair of equations (5.19.1) and (5.19.2). Since closed form solutions are obtained at every time instant and assuming the computation errors to be uniformly random, we perform smoothing on the time series of the computed parameters, to eliminate a large portion of the error.

The important aspect of this method of computation of the motion parameters is as follows:

(a) The solution is closed form, requiring no iteration or search.

(b) The constraints are derived from the observed tracking velocities and rotations. We do not need the optical flow measurements.

(c) Here the observables are, $(\theta, \theta', \dot{\theta}, \dot{\theta}', \ddot{\theta}, \ddot{\theta}')$. These can be measured quite accurately by analog measurement apparatus. This possibility forms a strong motivation for the tracking approach.

(d) The optical flow field, in our motion perception scheme is only used to disambiguate between the possible interpretations computed by the

tracking module. This is always possible since under extended periods of observation the optical flow field generated is compatible with one and only one interpretation.

5.5. Experiments

5.5.1. Stereo with parallel camera axes

We have carried out some preliminary experiments on artificial images to date. Assuming binocular vision and tracking we obtain A and γ from which we can recover the other parameters.

The experiments were performed under certain assumptions:

(a) The optical flow is known at each point.

(b) There are a reasonable number of points in the vicinity of the tracked point.

(c) The translational velocity parameter along the camera axis (i.e. Z axis) is small compared to the average scene depth.

The algorithm used to recover A and γ is as follows:

(a) Obtain possible stereo correspondences by epipolar constraints, i.e. the difference in the y values in the two camera's image frame has to lie within a certain value which we shall call the radius.

(b) Calculate the depth of the point by the correspondence.

(c) Throw away all correspondences which give extreme values of (depth of point - ρ) where ρ = depth of tracked point.

(d) Repeat step c until the number of points has been reduced to some threshold (typically the original number of points).

(e) Calculate the coefficients of A and γ in equation(5.3) for the remaining points, and apply the least squares method to obtain A and γ.

The experiments were performed for values of f (focal length) ranging from 35mm to 200mm, d (stereo baseline/2) ranging from 4 cm to 20 cm, θ (angle of rotation) varying between 2 degrees and 5 degrees and additive noise of up to 20 percent. We found that the algorithm was quite stable within these limits, recovering A and γ to within 10 - 25 percent accuracy. As the radius(distance between epipolar lines for correspondence) increased the error increased. Further, if steps (c) and (d) of the algorithm were not carried out, the errors were found to be much bigger, specially as the radius became large. The results are summarized in table 5.1. (Note that the values of V calculated from equation 5.9 are tabulated, together with γ).

The parameters relevant to table 5.1 are (the unit of length is one pixel width):

Focal length = 1000

stereo baseline = d = 1000

Rotation = (α,β,γ) = (0.0688 , 0.0229 , 0.0688)

Translation = (U , V , W) = (- 227 , 611 , 34)

Percentage of noise = 10

The algorithm works with large amounts of additive noise because most of the noise points get removed in step (c) of the algorithm. Other points whose depth is calculated to be very different from *p* also get removed, leaving points for which the A coefficient in equation (3,3) are quite similar which gives better results with least squares. The error is due to two factors:

(a) Discretization: This becomes specially important when the optical flow or the depths are small.

(b) Wrong correspondences: These may be reduced by using more elaborate statistical smoothing techniques in tandem with the parameter evaluation stage (e.g. the overall scheme can be a few iterations of a noise filtering step, then parameter hough transform followed by

radius	av. false match count	Y	V	error in *y* (%)	error inV (%)
0	0.76	0.0640	647	8	9
1	2.48	0.0634	539	7	12
2	3.83	0.0638	533	7	13
3	5.27	0.0636	526	7	14
5	7.83	0.0632	510	7	17
10	15.26	0.0645	485	8	21

Table 5,1 Measurements for tracking with stereo fusion

further pruning of the input data and so on).

5.5.2. Convergent Stereo

The geometric configuration in this case is depicted in figure 5.7. The simulation experiments were performed under the following assumptions:

(1) The precision of angular measurements is up to half a minute of arc. (i.e. the truncation error $\sim$ .0001 radians).

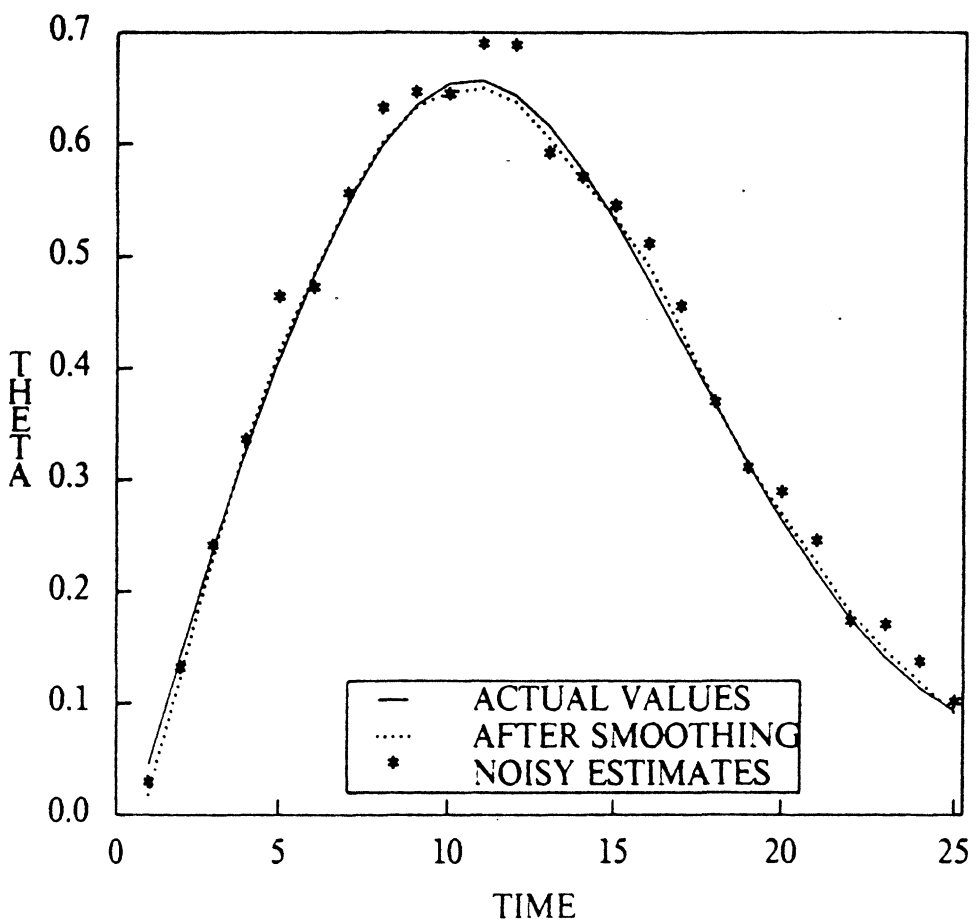


Figure 5.8 Time evolution of angular position

(2) The error in estimating the angular positions of the camera axes is random and follows a normal distribution with zero mean and standard deviation no more than five times the error due to truncation.

(3) The motion of the system is smooth. In particular the path of translation is piecewise linear in time, and the speed of translation changes very slowly (no acceleration). Furthermore, there is no precession or any other change in the rate or direction of rotation.

The motion stimuli were generated, synthetically, by applying exact rigid transformations (rotations and translations), using user specified parameters. The time progression was modeled by a sequence of small intervals (ticks). At every tick three additional rotations were generated to maintain tracking.

The output of the data generation program consists of the sequence of values of $\$$ and •', which are the angles made by the optical axes of the left and right camera respectively with the base line. Additionally, the rotation of the camera system about the baseline was also recorded *(w* of figure 5.1). All computations were done with respect to the left frame. The values of $\$$ and $\$'$ were artificially corrupted by random noise following a zero mean normal distribution, with standard deviation around 0.5°.

The steps in the computation were:

(i) The $\$$ values were smoothed over time to reduce the noise. An example of the effect of smoothing can be seen in figure 5.8, where the

actual values, the observed values and the smoothed values of θ are plotted. The scale for θ is in radians while that of 'time' is in ticks.

(ii) The depth values and its derivatives were computed from the smoothed curve $\theta(t)$, using the equations (5.15) and (5.16).

(iii) The ω_y component of the tracking rotation is simply the first derivative of $\theta(t)$ with its sign reversed. The other components, ω_x and ω_z were computed from $\omega(t)$ and $\theta(t)$, by $\omega_x = -\omega\sin\theta$ and $\omega_z = \omega\cos\theta$.

(iv) Finally the value of the body rotation is computed from equation (5.20). The remaining parameters of motion are obtained from equations (5.19),(5.17) and (5.18) by back substitution.

A typical set of values for the parameters used in the experiments is:

Stereo baseline length = 1.0

Initial value of $\theta = 100°$

Initial value of $\theta' = 77°$

Initial depth of tracked point = 18.62

Vector (unnormalized) specifying the rotation axis = (1, 2, 3)

The angle of rotation per time step = $2°$

The translation vector = (0.2, 0.3, 0.1)

The results for this set are plotted in figures 5.8 and 5.9. In figure 5.9 the actual value for the x component of the rotation and the two computed solutions (according to (5.19)) are shown. The error in the computed values of the rotation parameters were less than one percent. After back

substitution, the error in the values obtained for the translational parameters were found to be around 5%.

5.6. Summary & Conclusions

In this chapter a mathematical framework for active navigation, employing tracking, has been developed. The results reported here suggest that there is a better alternative to the "passive" technique for visual navigation [19, 52, 68]. This new approach is termed *Active Navigation*. The qualifier "active" is used because the mobile system is required to track an

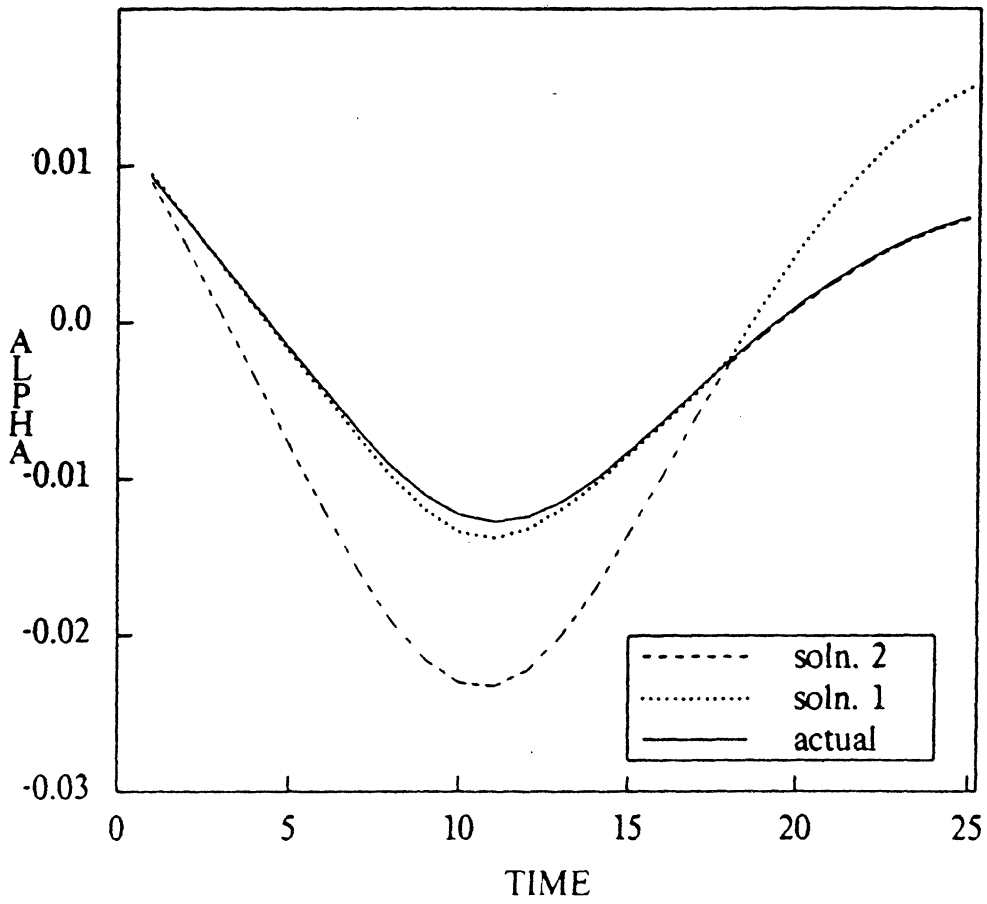


Figure 5.9 Time Evolution of rotation about x-axis

environmental feature.

The passive method has been unable to deliver practical algorithms for motion parameter estimation due to the fact that the constraint equations that arise are nonlinear and involve a fairly large number of unknowns. The computation in such cases is hampered by sensitivity to small errors and the need to have initial estimates of the solution in order to commence the search/iteration in the nonlinear parameter space [84]

The idea that tracking environmental points may be beneficial to navigation has previously been put forward by Cutting [22]. His analysis, however, is largely qualitative. A general analysis of the tracking geometry shows that the difficulties in motion parameter computation are alleviated under monocular imaging and largely removed for the binocular case. The problem with the binocular situation is that both motion as well as stereo correspondence is needed. Simulation experiments were conducted to examine the feasibility of this approach. The results are acceptable, when the the stereo fusional radius is is known. Therefore in itself, the stereo/motion approach cannot be recommended in practical cases due to accumulation of stereo and motion matching errors.

On the other hand, an analysis of a tracking system as in figure 5.1 shows that if the position, angular velocity and acceleration of the tracking motors can be measured over a period of time, then closed form solutions of the egomotion parameters are obtainable. In general, two solutions are

obtained at every time instant. However, an extended period of observation can disambiguate between these, since the two solution trajectories intersect at the correct solution (see figure 5.9).

Experimental simulation of the time evolution of of the solution space was conducted, using discretely generated motion data corrupted by noise that was as much as 20% of the rotation parameter value. The computational scheme proved to be quite robust with respect to these random noise fluctuations. The point is that the equations are stable enough, so that perturbations caused by noise are not overwhelming. Therefore, the correct solution trajectory can be recovered by temporal smoothing and interpolation.

Thus a strong case can be made for adopting this method for visual navigation, when the mobile system is undergoing steady motion. Even when the steady motion assumption holds only approximately (e.g. when there is a steady translational acceleration), the stability of the equations allow us to obtain reasonable estimates of the motion parameters. This suggests a cooperative scheme for the motion perception task. This involves using the closed form solutions obtained from the tracking constraints to be used as initial estimates in the monocular and binocular "flow" modules to refine the solution and compute structure of the observed surface. Such a scheme is outlined in the concluding chapter.

Chapter Six

Conclusion

6.1. Summary and Discussions

The purpose of this research was to analyze the problem of Rigid Body Motion Perception. The paradigm adopted for this study was a model for motion perception where the main task was viewed as the computation of sets of parameters that defined a hierarchy of abstraction levels. The parameters at any level can be thought of as succinct representations for the invariances that characterize that level. The computations performed at the proposed levels of the hierarchy span *Low Level* to *Intermediate Level* visual processing tasks.

The concept of spatial receptive field (SRF) of the parameters at any level of the hierarchy was introduced to capture the notion of the degree of abstraction realized by the parameters at that level. Thus at lower levels of the hierarchy, involving the computation of optical flow for instance, the SRF is small. In contrast, at the higher levels of abstraction, an example of

which have larger SRF. In the latter case, the SRF is global in the sense that it spans the part of the visual field that receives input from an entire moving surface.

The study was divided into three main stages. The first stage had to do with the measurement of image motion. It was demonstrated that *clustering* is a powerful tool in determining and structuring the image motion field. An orthogonal image decomposition scheme was also introduced to determine match tokens in time varying intensity images.

The second part dealt with the analysis of the algebraic constraints between the image motion field and the rigid body motion parameters. Here, the solution of two open problems were derived. These had to do with the upper bound on the *number of interpretations* of the optical flow field, which was proved to be three, and the *conditions* under which *unique interpretation* is possible. Regarding the latter question it was proved that the condition of ambiguity can be resolved by making observations at more than two time instants.

Finally, the analysis of the previous part were utilized to design algorithms for estimating the motion parameters from image motion, using the hough transform technique.

It was seen that nonlinearity and large dimensionality of the parameter space were two obstacles to the solution of the motion perception problem. In chapter five, alternative *Active*, strategies were proposed to tackle these

difficulties. The notion of tracking was introduced in analogy with the human smooth pursuit system. It was shown that under this active scheme the motion parameter estimation problem becomes simpler. In the case of egomotion, when the observer's motion is steady in space over the period of observation, it is possible to obtain *closed* form solutions for the egomotion parameters. In general, when the assumption of steady motion does not hold, the above parameter estimate degrades. However, the tracking solution is proposed, as one of the modules in a cooperative stereo-motion (see figure 6.1) system.

In this cooperative scheme, the motion and stereo correspondence is aided by the initial parameter estimate. Future research will determine the efficacy of this approach, although as far as motion and structure interpretation from optical flow is concerned it forms a vital link in the proposed hierarchical motion perception model (figure 1.1), since good initial estimates of the parameter values, as seen in chapter four, greatly simplifies the task of rigid body motion parameter computation.

6.2, Future Work

The research reported in this thesis, has opened up several avenues for further study. The task of motion perception is seemingly complex. The model proposed here is based on the belief that highly parallel hierarchies of simple local interactions is in principle upto the task. This belief is bolstered

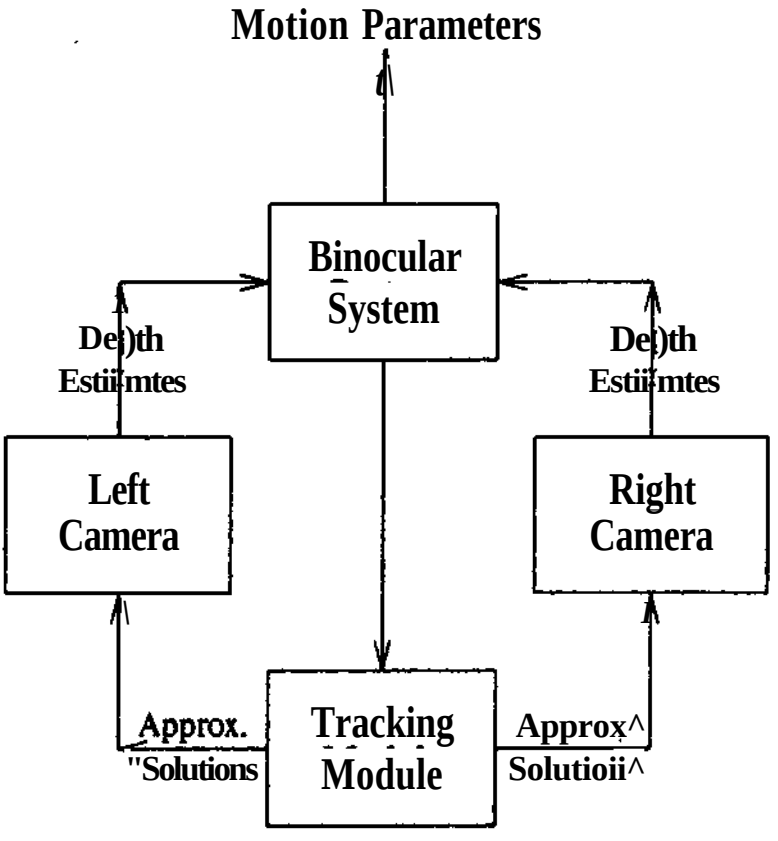


Figure 6.1 A cooperative model for motion perception

by the results demonstrated here and by the increasingly better understanding of the biological visual mechanisms evolving from research in psychology and the neural sciences.

Areas in which future research in computational studies in motion perception may be directed include:

(i) ***Correspondence using labeled interest points:*** **The orthogonal** decomposition operator provides a a natural way of labeling the selected interest points. These labels when used explicitly may reduce the

average number of match possibilities for the image motion algorithm.

(ii) *Cluster Analysis for multiple time frames*: The cluster based approach to image motion analysis is extendible to multiple time frames. This needs to be analyzed and experimentally evaluated.

(iii) *Limited Spatial indexing of the velocity space*: The clustering method is oblivious to the spatial coherence of the match data, only dealing with match vector or velocity values. It is easy to extend the clustering scheme to include some measure of spatial coherence like the principal components of the two dimensional spatial scatter. The details of such a strategy needs to be worked out and implemented.

(iv) *Alternative Clustering strategies*: A controlled probabilistic relaxation technique seems possible. Maximizing entropy or energy functionals are good candidates. This needs further study.

(v) *Sampling and quantization of the parameter space*: Uniform sampling and quantization need not be the only solution. The sampling properties of the motion parameter space needs to be studied to determine whether, for instance, random location and sizes of the cells lead to economy and efficiency without sharp decline in performance.

(vi) *Tracking target selection*: Aggregating many coherently moving features may help. The exact mechanism for doing this needs to be examined.

Bibliography

1. G. Adiv, "Determining 3-D Motion and Structure from optical flow generated from several moving objects", COINS-Tech. Rep. 84-07, Dept. of Computer Science, University of Massachusets at Amherst, July 1984.

2. J. Aloimonos and C. M. Brown, "Direct Processing of Curvilinear Sensor Motion from a Sequence of perspective Images", *Proc. IEEE Workshop on Computer Vision Representation and Control*, Annapolis, MD., 1984.

3. J. Aloimonos, A. Basu and C. M. Brown, "Contours, Shape, Motion", *DARPA IU workshop*, Miami, Florida, 1985.

4. J. Aloimonos and M. J. Swain, "Shape from Texture", *Proceedings of the International Joint Conf. on Artificial Intelligence*, **2**, (August 1985), 926-931.

5. J. Aloimonos, "One eye suffices : A computational model of monocular depth perception", Tech. Rep. 160, Dept. of Computer Science, Univ. Rochester, 1985.

6. **D. H. Ballard, "Parameter Networks: Towards a theory of Low Level Vision", 75, Computer Science Department, University of Rochester, 1981.**

7. **D. H. Ballard and C. Brown, *Computer Vision,* Prentice Hall, 1982.**

8- **D. H. Ballard and O. KimbaJl, "Rigid Body Motion From Depth and Optical Flow", *Computer Vision, Graphics & Image Processing,* 22, (1983), 95-115.**

9- **D. H. Ballard, "Parameter Nets", *Artificial Intelligence,* 22, (1984), 235-267.**

10. **D. H. Ballard, "Cortical connections and parallel processing: Structure and function", *The Behavioral and Brain Sciences, 9,* (1986), 67-120.**

11. **A. Bandopadhay and R. Dutta, "Measuring Image Motion in Dynamic Images", *IEEE Workshop on Motion Representation & Analysis,* Charleston, SC, May, 1986..**

12. **A. Bandopadhay, "Interest Points, Disparities and Correspondence", *DARPA Image Understanding Workshop,* 1985.**

13. A. Bandopadhay, "Constraints on the Computation of Rigid Motion Parameters **from Retinal Motion.", Tech. Rep. 168, Department of** Computer Science, The University of Rochester., October 1985.

14. S. T. Barnard and W. B. Thompson, "Disparity Analysis of Images", *IEEE Trans. PAMI,* 2, (1980), 333 - 340.

15. H. G. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images", in *Computer Vision Systems*, A. R. Haanson and E. M. Riseman, (eds.), Academic Press, 1978.

16. C. M. Brown, "An iterative improvement algorithm for coherent codes", *Optics Comm.*, June 1980.

17. C. M. Brown, M. B. Curtiss and D. B. Sher, "Advanced Hough Transform Implementations", *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 1983, 1081.

18. C. M. Brown, "Hierarchical Cache Accumulators For Sequential Mode Estimation", Tech. Rep. 125, Computer Science Department, University of Rochester, July 1983.

19. A. Bruss and B. K. P. Horn, "Passive Navigation", *Computer Vision, Graphics & Image Processing*, **21**, (1983), 3-20.

20. J. F. Canny, "Finding Edges and Lines in Images", MIT-AI-Tech. Rep. 720, AI Laboratory, M.I.T., June 1983.

21. H. S. M. Coxeter, *Introduction to Geometry*, John Wiley & Sons.

22. J. E. Cutting, "Motion Parallax and Visual Flow: How to Determine Direction of Locomotion", (Paper presented at the 4th meeting of the International Society for Ecological Psychology, Hartford, CT.,1982), Dept. of Psychology, Cornell University, 1982.

23. P. E. Danielsson, "Rotation-invariant linear operators with Directional Response", *Proc. 5th Int. Conf. on Pattern Recognition*, December 1980.

24. L. S. Davis, Z. Wu and H. Sun, "Contour Based Motion Estimation", *Computer Vision, Graphics & Image Processing*, **23**, (1983), 313-326.

25. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

26. D. C. V. Essen and J. H. R. Maunsell, "Hierarchical organization and functional streams in the visual cortex", *Trends in Neurosciences*, **6**, (1983), 370-375.

27. J. Q. Fang and T. S. Huang, "A corner finding algorithm for Image Analysis", *Proc. AAAI-82*, Pittsburgh, Pennsylvania, August 1982, 46-49.

28. J. Q. Fang and T. S. Huang, "Solving three dimensional small-rotation motion equations: Uniqueness, algorithms, and numerical results", *Computer Vision, Graphics & Image Processing*, **26**, (1984), 183-206.

29. J. Q. Fang and T. S. Huang, "Some Experiments on Estimating the 3D Motion Parameters of a Rigid body from two Consecutive Image Frames", *IEEE Trans. PAMI*, **6(5)**, (1984), 545-554.

30. J. A. Feldman, "Connectionist Models and Parallelism in High Level Vision", Tech. Rep. 146, Department of Computer Science,

University of Rochester, January, 1985.

31. J. A. Feldman and D. H. Ballard, "Connectionist Models and Their Properties", *Cognitive Science*, **6**, 205 - 254.

32. J. A. Feldman, "Four Frames Suffice", Tech. Rep. 99, Computer Science Department, University of Rochester.

33. C. L. Fennema and W. B. Thompson, "Velocity Determination in Scenes Containing Several Moving Objects", *Computer Vision, Graphics & Image Processing*, **9**, (1979), 301-315.

34. D. J. Fleet and A. D. Jepson, "A Cascaded Filter Approach to the The Construction of Velocity Selective Mechanisms", RBCV-Tech. Rep.-84-6, Department of Computer Science, University of Toronto, December 1984.

35. W. Frei and C. C. Chen, "Fast boundery detection: A Generization and a New Algorithm", *IEEE Trans. Comput.*, **26**, (1977), 988 - 998.

36. J. J. Gibson, *Perception of The Visual World*, Riverside Press Cambridge, Massachusets, 1950.

37. T. F. Gonzales, "Clustering to Minimize the Maximum Intercluster Distance", *Theoretical Computer Science*, **38**, (1985), 293-306.

38. R. M. Haralick and J. S. Lee, "The Facet Approach to Optical Flow", *Proc. Image Understanding Workshop*, Arlington, Virginia, June, 1983, 84 - 93.

39. H. L. F. V. Helmholtz, *Treatise of Physiological Optics*, NY Dover Publications, 1925.

40. E. C. Hildreth, "Computations Underlying the Measurement of Visual Motion", *Artificial Intelligence*, **23**, (1984), 309-354.

41. D. D. Hoffman and B. E. Flinchbaugh, "The Interpretation of Biological Motion", *Biol. Cybernetics*, **42**, (1982), 195-204.

42. D. D. Hoffman, "Inferring local surface orientation from motion fields", *J. Opt. Soc. Am.*, **72(7)**, (July 1982), 888-892.

43. D. D. Hoffman and B. M. Bennett, "Inferring the relative three dimensional positions of two moving points", *J. Opt. Soc. Am. A*, **2(2)**, (February 1985), 350-353.

44. B. K. P. Horn and B. G. Schunck., "Determining Optical Flow", *Artificial Intelligence*, **17**, (1981), 185-204.

45. T. S. Huang and R. Y. Tsai, "Image Sequence Analysis: Motion Estimation", in *Image Sequence Analysis*, T. S. Huang, (ed.), Springer-Verlag, 1981, ch. 1.

46. D. H. Hubel and T. N. Wiesel, "Receptive field and Functional Archiecture of Monkey Striate Cortex", *The Journal of Psychology*, **195**, 2 (Nov 1968), 215 - 244.

47. D. H. Hubel and T. N. Wiesel, "Brain Mechanisms of Vision", *Scientific American*, Sept. 1979.

48. R. Jain, "An approach for the Direct Computation of the Focus of Expansion", *PRIP*, 1982, 262-268.

49. R. Jain, *'Direct Computation of the Focus of Expansion", *IEEE Trans. PAMI*, 5(1), (January 1983), 58-63.

50. Jenkin, "The stereopsis of time-varying images", Tech. Rep. RCBV-Tech. Rep.-84-3, University of Toronto, Comp. Sci Dept, 1984.

51. L. Kitchen and A. Rosenfeld, "Gray - Level Corner Detection ", Tech. Rep. 887, Computer Science Center, University of Maryland, 1980.

52. H. C. LongueirHiggins and K. Prazdny, "The interpretation of a moving retinal image", *Proc. Royal Soc. London*, B 208, (1980), 385 - 397.

53. H. C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections", *Nature*, 283, (September 1981), 133-135.

54. **J.** S. Lund, "Intrinsic organization of the primate visual cortex, area **17,** as seen in Golgi preparations", in *The organization of the cerebral cortex,* F. O. Schmitt, F. G. Worden, G. Adelman and S. G. Dennis, (eds.), MIT Press, 1981.

55. D. Marr and E. Hildreth, "Theory of Edge Dection", *Proc. Royal Soc. London. Series B.,* **207,** (1980), 187-217.

56. D. Marr and *S.* Ullman, "Directional Selectivity and its Use in Early Visual Processing", *Proc. Royal Soc. London, Ser. B,* 211, (1981), 151-180.

57. D. Marr, *VISION,* W.H. Freeman, San Francisco, 1982.

58. H. P. Moravec, "Towards Automatic Visual obstacle avoidance", *Proc. 5th International Joint Conf. on Artificial Intelligence.* , 1977, 584.

59. J. A. Movshon, "Analysis of visual motion", in *Proc. Conference on Vision, Brain and Cooperative Computation,* M. Arbib and A. Hanson, (eds.), MIT Press, 1983.

60. H. H. Nagel and B. Neumann, "On 3-D Reconstruction from Two Perspective Views[11], *International Joint Conf. on Artificial Intelligence,* 1981, 661-663.

61. H. H. Nagel, "Overview on Image Sequence Analysis", in *Image Sequence Processing & Dynamic Scene Analysis,* T. S. Huang, (ed.), 1983.

62. H. Nagel, "Displacement Vectors Derived from Second order Intensity Variations **in** Image Sequences", *Computer Vision, Graphics & Image Processing,* **21,** (1983), **85** - 117.

63. K. Nakayama, "Biological Image Motion Processing: A Review", *Vision Research,* **25(5),** (1985), 625-660.

64. S. Negahdaxipour and B. K. P. Horn, "Determinig 3-D motion of planar objects from image brightness patterns", *Proceedings of the* ***International Joint Conf. on Artificial Intelligence,*** **2, (August 1985), 898-901.**

65. R. Paquin and E. Dubois, "A Spatio-Temporal Gradient Method for Estimating the Displacement Field in Time Varying Imagery'*, ***Computer Vision, Graphics & Image Processing,*** **21, (1983), 205-221.**

66. J. M. Prager and M. A. Arbib, "Computing The Optic Flow: The MATCH Algorithm and Prediction", *Computer Vision, Graphics & Image Processing,* **21,** (1983), 271-304..

67. K. Prazdny, [C4]Egomotion and Relative Depth Map from Optical Flow", *Biol. Cybernetics,* 36, (1980), 87-102.

68. K. Prazdny, "Determining the instantaneous direction of motion from opticl flow generated by curvilinearly moving observer", *Computer* ***Vision, Graphics & Image Processing,*** **17, (1981), 94 - 97.**

69o K. Prazdny, "Detection of Binocular Disparities", *Biol. Cybern.,* 52, (1985), 93-99.

70. J. H. Rieger and D. T. Lawton, "Determining the Instantaneous Axis of Translation from Optic Flow Generated by Arbitrary Sensor Motion", *COINS tech. report 88-1.,* January 1983.

71. J. W. Roach and J. K. Aggarwal, "Determining the Movement of Objects from a Sequence of Images", *IEEE Trans. PAMI*, **2** (6), (November 1980), 54-562.

72. D. A. Robinson, "The mechanics of human saccadic eye movement", *Journal of Physiology*, **174**, (1964), 245-264.

73. D. A. Robinson, "The mechanics of human smooth pursuit eye movements", *Journal of Physiology*, **180**, (1965), 569-591.

74. H. Sakata, H. Shibutani and K. Kawano, "Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey", *Journal of Neurophysiology*, **43**, (1980), 1654-1672.

75. M. I. Shamos, "Geometry and Statistics: Problems at the Interface", in *Algorithms and Complexity: New Directions and Recent Results*, J. F. Traub, (ed.), Academic Press, 1976, 251-280.

76. M. Sollin, "An Algorithm for finding Minimum Spanning Tree (attributed to Sollin)", in *Introduction to The Design and Analysis of Algorithms*, vol. Section 5.5, S. E. Goodman and S. T. Hedetniemi, (eds.), McGraw-Hill, New York, 1977.

77. M. J. Steinbach, "Pursuing the perceptual rather than the retinal stimulus", *Vision Res.*, **16**, (1976), 1371-1376.

78. K. A. Stevens, "Computation of Locally Parallel Structure", *Biol. Cybern.*, **29**, (1978), 19-28.

79. K. Sugihara and N. Sugie, "Recovery of Rigid Structure from Orthographically Projected optical flow", *Computer Vision, Graphics & Image Processing*, 27, (1984), 309-320.

80. C. E. Thorpe, "An Analysis of Interest Operators for FIDO", CMU-RI-Tech. Rep.-83-19, The Robotics Institute, Carnegie Mellon University, December 1983.

81. R. Y. Tsai and T. S. Huang, "Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch", *IEEE Trans. A.S.S.P.,* **ASSP-29,** (December 1981), 1147-1152.

82. R. Y. Tsai, T. S. Huang and W. Zhu, "Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch II : Singular Value Decomposition", *IEEE Trans. A.S.S.P.,* **ASSP-30,** (August 1982), 525-534.

83. R. Y. Tsai and T. S. Huang, "Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch III: Finite Point Correspondences and the Three-View Problem", *Proc. IEEE Conf. ASSP, Paris ,* May 1982.

84. R. Y. Tsai and T. S. Huang, "Uniqueness and Estimation of Three Dimensional Motion Parameters of Rigid Objects with Curved Surfaces", *IEEE, Trans. PAMI,* 6, (January 1984), 13-27.

85. S. Ullman, "The Interpretation of Visual Motion", *Ph.D. Thesis*, 1977.

86. S. Ullman, "The Interpretation of Structure from Motion", *Proc. R. Soc. Lond. (B)*, **B 203**, (1979), 405-426.

87. S. Ullman and E. Hildreth, "The Measurement of Visual Motion", in *Physical and Biological Processing of Images (Proc. Int. Symp. Rank Prize Funds. London)*, O. J. Braddick and A. C. Sleigh, (eds.), Springer-Verlag, September 1982, 154 - 176.

88. S. Ullman, "Maximising rigidity: The incremental recovery of 3D structure from rigid and nonrigid motion", *Perception*, **13**, (1984), 255-274.

89. A. M. Waxman and S. Ullman, "Surface structure and 3D Motion from Image Flow: A Kinematic Analysis", CAR-Tech. Rep.-24, Center for Automation Research, University of Maryland., October 1983.

90. A. M. Waxman and K. Wohn, "Contour evolution, neighbourhood deformation and global image flow:planar surfaces in motion", CAR-Tech. Rep.-74, Center for Automation Research, University of Maryland, April, 1984.

91. A. M. Waxman and J. H. Duncan, "Binocular Image Flows: Steps toward Stereo - Motio Fusion", CAR-Tech. Rep.-119, Computer Vision Laboratory, University of Maryland., May 1985.

92. J. Webb and J. K. Aggarwal, "Structure from motion of rigid aj jointed objects", *Artificial Intelligence,* 19, (1982), 107-130.

93. G. Westheimer and S. M. Mckee, "Visual acuity in the presence retinal motion", /. *Opt Soc. Am.,* 66, (1975), 847-850.

Appendix A

Uniqueness of Rigid Motion Parameters

Consider a point *P* in space whose coordinates are *[X,Y$_f$Z)* with respect to a fixed inertial frame XYZ- The image of this point is *p* =(z,y) whose coordinates are given with respect to a xy frame located on the image plane. The relation between the world point P and the image point p is given by

$$(z,y) = (f \wedge \quad) \qquad (1)$$

where T is the *focal length* of the imaging system. This is assumed to be unity in the following analysis.

Now if a rigid surface moves with a translational velocity V_T =((7,y,w) and a rotational velocity fi =(a,0,7). Then, from kinematics, the three dimensional velocity of any point on the surface can be written as

where 't' is the time variable and ^{4}x' denotes vector product.

In differential motion case the image motion or optical flow is denoted by $(u_t v)$ $=(\sim^{\partial}\!/_{\partial t}, \sim^{\partial}\!/_{\partial t})_{\#}$ Differentiating equation (1) and substituting from equation (2) we have the following relations

$$u = \frac{U - xW}{Z} - \alpha xy + \beta(x^2 + 1) - \gamma y \qquad \text{(iii.a)}$$

$$v = \frac{V - yW}{Z} - \alpha(y^2 + 1) + \beta xy + \gamma x \qquad \text{(iii.b)}$$

Eliminating the unknown depth variable from the above we get

$$\frac{u + \alpha xy - \beta(x^2 + 1) + \gamma y}{v + \alpha(y^2 + 1) - \beta xy - \gamma x} = \frac{U - xW}{V - yW} \qquad \text{(iv)}$$

The above equation describes the constraint imposed by the measured value of optical flow (u,v), at an image point (x,y), on the six motion parameters $(U,V,W,\alpha,\beta,\gamma)$.

Proposition I. *Given the rotation parameters the translation parameters can be uniquely determined from the optical flow field*

Proof: First we define the function $\mu(x,y)$ where,

$$\mu = \frac{u + \alpha xy - \beta(x^2 + 1) + \gamma y}{v + \alpha(y^2 + 1) - \beta xy - \gamma x} \qquad \text{(iv)}$$

Now we analyse the following cases:

Case 1: If $\mu = constant$ then from equation (iv) we have $W = 0$. In this case we can only obtain the ratio $\frac{U}{V}$ from the optical flow field.

Case 2: If $\mu \neq constant$ then there are two image points where μ is different. In which case we can solve the resultant set of two linear equations, obtained from (iv), to get $x_0 = \frac{U}{W}$ and $y_0 = \frac{V}{W}$.

Proposition II. *Given the translation parameters the rotation parameters can be uniquely determined from optical flow.*

Proof: Here the values of x_0 and y_0 are known. The expression for optical flow is,

$$u = (x_0 - x)\phi - \alpha xy + \beta(x^2 + 1) - \gamma y$$

$$v = (y_0 - y)\phi - \alpha(y^2 + 1) + \beta xy + \gamma x$$

Where (α,β,γ) are the rotation parameters and $\phi = \frac{W}{Z}$ is the reciprocal of the scaled depth function. If possible let there be another surface moving with the same translation but different rotation parameters, but generating the same otical flow. Thus we have,

$$u = (x_0 - x)\phi' - \alpha' xy + \beta'(x^2 + 1) - \gamma' y$$

$$v = (y_0 - y)\phi' - \alpha'(y^2 + 1) + \beta' xy + \gamma' x$$

Now from the above sets of equations by subtracting appropriately we get,

$$0 = (x_0 - x)(\phi - \phi') - \Delta\alpha xy + \Delta\beta(x^2 + 1) - \Delta\gamma y \qquad \text{(v.a)}$$

$$0 = (y_0 - y)(\phi - \phi') - \Delta\alpha(y^2 + 1) + \Delta\beta xy + \Delta\gamma x \qquad \text{(v.b)}$$

where $\Delta\alpha = \alpha - \alpha'$, $\Delta\beta = \beta - \beta'$ and $\Delta\gamma = \gamma - \gamma'$. Eliminating $(\phi - \phi')$ from the above we have,

$$(\Delta\alpha x_0 + \Delta\beta y_0) - x(\Delta\gamma x_0 + \Delta\alpha) - y(\Delta\gamma y_0 + \Delta\beta) + x^2(\Delta\beta y_0 + \Delta\gamma)$$

$$+ y^2(\Delta\alpha x_0 + \Delta\gamma) - xy(\Delta\beta x_0 + \Delta\alpha y_0) = 0 \qquad \text{(vi)}$$

Since the above equation is valid everywhere in the image,

$$\Delta\alpha x_0 + \Delta\beta y_0 = 0 \qquad \Delta\alpha y_0 + \Delta\beta x_0 = 0$$

$$\Delta\gamma x_0 + \Delta\alpha = 0 \qquad \Delta\beta y_0 + \Delta\gamma = 0$$

$$\Delta\gamma y_0 + \Delta\beta = 0 \qquad \Delta\alpha x_0 + \Delta\gamma = 0$$

From the above we obtain,

$$\Delta\alpha = 0 \qquad \Delta\beta = 0 \qquad \Delta\gamma = 0$$

This means that $\alpha = \alpha'$, $\beta = \beta'$ and $\gamma = \gamma'$ and therefore, *the rotation parameters are uniquely determined* when the translation parameters are known.

Proposition III *If the structure of a Rigidly moving surface is known, then the parameters describing its motion is uniquely determined.*

Proof: Knowing structure means that we have the depth values available up to some scale factor. Thus in equation (iii) the value 'Z' is no longer an unknown. The unknown scale factor is lumped with the translation parameters. Now proceeding in a manner analogous to the previous proof we have,

$$\frac{1}{Z}(\Delta U - x\Delta W) = \Delta\alpha xy - \Delta\beta(x^2 + 1) + \Delta\gamma y \qquad \text{(vii.a)}$$

$$\frac{1}{Z}(\Delta V - y\Delta W) = \Delta\alpha(y^2 + 1) - \Delta\beta xy - \Delta\gamma x \qquad \text{(vii.b)}$$

Eliminating $\frac{1}{Z}$ we have,

$$(\Delta\alpha\Delta U + \Delta\beta\Delta V) - x(\Delta\gamma\Delta U + \Delta\alpha\Delta W) - y(\Delta\beta\Delta W + \Delta\gamma\Delta V)$$
$$+ x^2(\Delta\gamma\Delta W + \Delta\beta\Delta V) + y^2(\Delta\gamma\Delta W + \Delta\alpha\Delta U) - xy(\Delta\alpha\Delta V + \Delta\beta\Delta U)$$

Since the above equation must be valid all over the image plane, the following relations hold:

$$\Delta\alpha\Delta U + \Delta\beta\Delta V = 0 \qquad \Delta\alpha\Delta W + \Delta\gamma\Delta U = 0 \qquad \Delta\beta\Delta W + \Delta\gamma\Delta V = 0$$

$$\Delta\alpha\Delta V + \Delta\beta\Delta U = 0 \qquad \Delta\beta\Delta V + \Delta\gamma\Delta W = 0 \qquad \Delta\alpha\Delta U + \Delta\gamma\Delta W = 0$$

From eqn. (vii) and the above relations we have,

$$\Delta U = \Delta V = \Delta W = \Delta\alpha = \Delta\beta = \Delta\gamma = 0$$

Therefore, once the structure is known for a rigidly moving surface, its translation (up to a scale factor) and its rotation is determined uniquely from the optical flow generated by the motion.

Appendix B

Representation of surface shape

In computer vision, the terms *surface orientation map* and *shape* are sometimes used interchangably. The following is an attempt to explain the basis of this usage. The cases of *Perspective* as well as *Orthographic* projections are considered. Shape information obtainable from a surface orientation map in image coordinates is also explored.

Representations for surface orientation

A direction in three space is specified by two independent parameters.

A. (Latitude, Longitude): The coordinates are denoted by (θ,ϕ) where $0 \leq \theta < \pi$, $0 \leq \phi < \pi$.

B. Coordinates on the *gaussian* (or unit radius) sphere. If the coordinates are (l,m,n) then $l^2 + m^2 + n^2 = 1$.

C. (slant , tilt): Slant is the tangent of ther latitude angle (or $\tan\theta$) while tilt is the longitude angle. The symbolic notation is (σ,τ).

D. (Gradient): If the depth is expressed in the form $Z = f(X,Y)$, then it is the level surface $F(X,Y,Z) = 0$, where

$$F\{X,Y,Z) = f(X,Y) - Z$$

The gradient of $F\{-)$, i.e. $\{\%,\%,-1)$ gives the orientation of the surface (in the direction of increasing $F(-)$). The gradient notation is written as $\{p_9q)_f$ where (p,g) $=(\frac{\partial f}{\partial X}, \frac{\partial f}{\partial Y})$.

Relationship among the surface normal representations:

$$\sqrt{p^2 + q^2} = \tan\theta = \sigma$$

$$\frac{q}{p} = \tan\phi = \tan\tau$$

$$(l,m,n) = (\frac{p}{g}, \frac{q}{g}, \frac{-1}{g}), \qquad g = \sqrt{p^2 + q^2 + 1}$$

Shape under Perspective Projection

In the case of perspective projection the relation between a world point $(X>Y,Z)$ and its projection (*,y) in the image plane is given by

where F is the focal length of the imaging system.

The surface is represented in the world frame by the functional form $Z(X,Y)$. It is assumed that the surface can also be represented (at least locally) by the function $z\{x,y)$ in image coordinates. Here the relation between the surface normals $(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y})$ corresponding to an image point $lx,y)$ and the partial derivatives of $z\{x>y)$ are saught.

A- *Relationship between surface gradients in image and world coordinates.* Now a small displacement (Sx_t6y) in the image plane corresponds to a

displacement $(\delta X, \delta Y, \delta Z)$ in the world frame, along the surface $Z(X,Y)$. From equation (i) we get the relation

$$\delta X = \frac{\delta x Z + x \delta Z}{F} \tag{ii.a}$$

$$\delta Y = \frac{\delta y Z + y \delta Z}{F} \tag{ii.b}$$

Furthermore the following identity holds

$$Z(X + \delta X, Y + \delta Y) = z(x + \delta x, y + \delta y) \tag{iii}$$

Using the Taylor series expansion of the above

$$Z(X + \delta X, Y + \delta Y) = Z(X,Y) + \delta X \frac{\partial Z}{\partial X} + \delta Y \frac{\partial Z}{\partial Y} + (\textit{higher order terms}) \tag{iv.a}$$

$$z(x + \delta x, y + \delta y) = Z(x,y) + \delta x \frac{\partial Z}{\partial x} + \delta y \frac{\partial Z}{\partial y} + (\textit{higher order terms}) \tag{iv.b}$$

Neglecting the higher order terms in equation (iv) and substituting for δX and δY from equation (ii) in equation (iv.a)

$$Z(X + \delta X, Y + \delta Y) - Z(X,Y) = \delta Z = \frac{1}{F}(\delta x Z + x \delta Z)\frac{\partial Z}{\partial X} + \frac{1}{F}(\delta y Z + y \delta Z)\frac{\partial Z}{\partial Y}$$

or

$$\delta Z\left(F - x\frac{\partial Z}{\partial X} - y\frac{\partial Z}{\partial Y}\right) = Z \delta x \frac{\partial Z}{\partial X} + Z \delta y \frac{\partial Z}{\partial Y} \tag{v}$$

Recall now that

$$Z(X + \delta X, Y + \delta Y) - Z(X,Y) = z(x + \delta x, y + \delta y) - z(x,y)$$

Therefore combining equations (iii), (iv) and (v)

$$\delta x \frac{Z}{F - x\frac{\partial Z}{\partial X} - \frac{\partial Z}{\partial Y}} \frac{\partial Z}{\partial X} + \delta y \frac{Z}{F - x\frac{\partial Z}{\partial X} - \frac{\partial Z}{\partial Y}} \frac{\partial Z}{\partial Y} = \delta x \frac{\partial z}{\partial x} + \delta y \frac{\partial z}{\partial y} \tag{vi}$$

Since δx and δy are independent of each other we have

$$\frac{\partial z}{\partial x} = \frac{Z\frac{\partial}{ax}}{F\text{-}nU.\text{-}\frac{\partial Z}{dY}} \qquad i - \backslash \quad (vu,a)$$

$$\frac{\partial z}{\partial y} = \frac{Z^{**}\text{-}}{F - s\frac{\partial Z}{\partial X} - \frac{dz}{\partial Y}} \qquad (vu.b)$$

B. ***What Shape means*** **Consider the shape information available from the field of surface normals indexed by the image coordinates. Making the appropriate substitutions from equations (vii) in equation (iv.b) we have:**

$$\frac{zlz + Sx\ ,\ y + Sy) \sim \text{«}}{Z\{*JV)} \quad {}^{P}\ \frac{dX}{p - z\frac{92}{8X''} - \frac{8Z}{dY}} \quad {}^{c}\ \frac{dY}{r - z\frac{dZ}{dX''} - \frac{dZ}{dY}}$$

Thus the following statement can be made:

> ***Under perspective projection, when the field of surface normals is available, indexed by image coordinates, then the image centered depth function can be computed upto a dilation factor.***

Lemma I. If the surface Z is represented by an ***algebraic function Z(X,Y)*** **and furthermore if the function** $z(x_9y)$ **denotes the same surface in terms of the image coordinates (x,y), then the** ***tilt*** **function** $r(x,y)$ **is given by**

$$r = \frac{\frac{dZ}{dX}}{\frac{dZ}{dY}} = \frac{\frac{dz}{dx}}{\frac{dz}{dy}}$$

Proof: Since $Z\{X_9Y)$ **is an algebraic function, by definition it can be expressed implicitly by the polynomial equation** $F(X_yY_jZ) = 0$**. We can write**

$F(\cdot)$ as

$$\sum_{i=0}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} c_{ijk} X^i Y^j Z^k = 0 \qquad \text{(viii)}$$

where the c_{ijk}'s are real constants and L, M, N are finite positive integers. By using the implicit function theorem we get

$$\tau = \frac{\dfrac{\partial Z}{\partial Y}}{\dfrac{\partial Z}{\partial X}} = \frac{-\dfrac{F_Y}{F_Z}}{-\dfrac{F_X}{F_Z}} = \frac{F_Y}{F_X}$$

where F_X, F_Y, F_Z denote the partial derivative of $F(\cdot)$ with respect to X, Y and Z. Therefore we have from equation (viii):

$$\tau = \frac{\sum_{i=0}^{L}\sum_{j=1}^{M}\sum_{k=0}^{N} j c_{ijk} X^i Y^{j-1} Z^k}{\sum_{i=1}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} i c_{ijk} X^{i-1} Y^j Z^k} \qquad \text{(ix)}$$

Observe now that we can obtain an implicit representation for the depth in terms of the image coordinates (x,y) from equation (viii) by substituting for X and Y in accordance with $x = \frac{X}{Z}$ and $y = \frac{Y}{Z}$ (where the focal length is assumed to be 1). Thus we obtain the representation $G(x,y,z) = 0$ or

$$\sum_{i=0}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} c_{ijk} x^i y^j z^{i+j+k} = 0 \qquad \text{(x)}$$

Again by the implicit function theorem we have

$$\frac{\dfrac{\partial z}{\partial y}}{\dfrac{\partial z}{\partial x}} = \frac{-\dfrac{G_y}{G_z}}{-\dfrac{G_x}{G_z}} = \frac{G_y}{G_x} = \frac{\sum_{i=0}^{L}\sum_{j=1}^{M}\sum_{k=0}^{N} j c_{ijk} x^i y^{j-1} z^{i+j+k}}{\sum_{i=1}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} i c_{ijk} x^{i-1} y^j z^{i+j+k}}$$

or

$$\frac{\frac{\partial z}{\partial y}}{\frac{\partial z}{\partial x}} = \frac{\sum_{i=0}^{L}\sum_{j=1}^{M}\sum_{k=0}^{N} j c_{ijk} x^i y^{j-1} z^{i+j+k-1}}{\sum_{i=1}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} i c_{ijk} x^{i-1} y^j z^{i+j+k-1}} \qquad \text{(xi)}$$

Consider now, equation (ix) and substitute $x = xz$ and $y = yz$

$$\frac{\frac{\partial Z}{\partial Y}}{\frac{\partial Z}{\partial X}} = \frac{\sum_{i=0}^{L}\sum_{j=1}^{M}\sum_{k=0}^{N} j c_{ijk} x^i y^{j-1} z^{i+j+k-1}}{\sum_{i=1}^{L}\sum_{j=0}^{M}\sum_{k=0}^{N} i c_{ijk} x^{i-1} y^j z^{i+j+k-1}} \qquad \text{(xii)}$$

But the right hand sides of the equations (xi) and (xii) are identical. This means,

$$\tau = \frac{\frac{\partial Z}{\partial Y}}{\frac{\partial Z}{\partial X}} = \frac{\frac{\partial z}{\partial y}}{\frac{\partial z}{\partial x}}$$

which concludes the proof of the lemma.

Shape under Orthographic Projection:

Under orthography the image coordinates of a point are equal to the corresponding three dimensional coordinates, or

$$(x, y) = (X, Y)$$

Thus

$$\left(\frac{\partial Z}{\partial x}, \frac{\partial Z}{\partial y}\right) = \left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right)$$

Now observe from equation (iv.a) that when the surface normals are known at an image point (x, y), then the depth difference between this point and neighbouring image points are known:

$$Z(X+SX,\ Y+SY)-\ Z\{X_tY)\ \ ^SX^= +\ BY^- +\ (higher\ order\ terms)$$

Thus we can state the following:

When a map of surface normals is available under orthography, the depth function can be computed upto a constant additive term.