Preface

This book is based on my doctoral dissertation completed in 1986 at Computer Vision Laboratory, University of Maryland. The original thesis has been revised and updated in many respects. Also some new results have been added.

A changing scene produces a changing image or *visual motion* on the eye's retina. The human visual system is able to recover useful three-dimensional information about the scene from this two-dimensional visual motion. This thesis is a study of this phenomenon from an information processing point of view. A *computational theory* is formulated for recovering the scene from monocular visual motion. This formulation deals with determining the local geometry and the rigid body motion of surfaces from spatio-temporal parameters of visual motion. Based on this formulation, a *computational approach* is presented. A notable characteristic of this approach is a uniform representation scheme and a unified algorithm which is flexible and extensible.

Visual motion, also referred to as *optical flow* in the literature, has been the topic of intense research activity in the last few years. An important outcome of this study is combining a large body of previous work to yield a coherent theoretical framework and a unified computational approach. Many of the results originally obtained by other researchers and by myself in the early stages of my research in this area have been rederived, sometimes by a simpler method, in the new framework. Thus, this thesis provides a theoretical and computational framework for future research on visual motion, both in human vision and machine vision areas.

I have made an attempt to make this book somewhat self-contained and suitable for reading by a non-specialist. The basic concepts behind the approach taken here and a summary of the results in this thesis are given in Chapter 1. The background material necessary to understand the later chapters is provided through a broad introduction to the field and review of previous literature in Chapter 2. Mathematical details of many results and implementation details of many computational algorithms are excluded from the main body of the book. They are included in the appendices as they are useful in computer vision applications.

Stony Brook, N.Y.   M. S.
January, 1988.

Acknowledgements

It is a delight to thank here those people who made this thesis possible, and this period of my life enjoyable and exciting.

Dr. Allen Waxman, who has influenced every aspect of this work, both in spirit and in content. His critical and insightful comments at every stage of this research have been invaluable to me.

Dr. Larry Davis, who has been a marvelous adviser to me all through my graduate studies. He and Dr. Waxman have given me immense freedom and have been extremely encouraging and supportive of the paths I have taken.

Dr. Azriel Rosenfeld, for providing me the unique opportunity to pursue my research in the Computer Vision Laboratory. His comments on this research has contributed greatly towards improving it. It is a pleasure to express my deep gratitude to him.

Dr. Behrooz Kamgar-Parsi, Prof. Ken-ichi Kanatani (of Gunma University, Japan), and the anonymous referees of my research papers have offered their valuable comments at various stages of this work. It is a delight to express my thanks to them all.

I express my heartfelt thanks to my friends in Maryland and in Boston for good friendship and great times! Shah Ashiquzzaman, Roger Eastman, David Harwood, Venugopal Iyengar, Stephen Omohundro, Raja Sekar, Yogendra Simha, Babu Srinivasan, Kambhampati Subbarao, Kwangyoen Wohn, and many more. I have learnt a great deal from these people and have enjoyed their friendship immensely.

A significant part of this research was done at the Computer Vision Laboratory, University of Maryland. The staff of the laboratory has been extremely helpful to me in all administrative matters. My sincere thanks to all members of the staff. The remaining part of this research was done at Thinking Machines Corporation, Cambridge, Massachusetts, where excellent

computing and office facilities were made available to me. I thank the concerned people for this and for their encouragement in my graduate research.

This research was funded by the Defense Advanced Research Projects Agency and the United States Army Night Vision and Electro-Optics Laboratory under Contract DAAK70-83-K-0018 (DARPA Order 3206). I gratefully acknowledge this support.

Most importantly, my mother Sharada and father Subbarao, who are my model of ideal parents. My sisters Krishnaveni and Manjula, and brothers Sreesha and Jagannatha, for their love and their support all along.

.*vs*20

CHAPTER 1

Introduction

> *‘‘One of the principal objects of theoretical research in any department of knowledge is to find the point of view from which the subject appears in its greatest simplicity’’.* -J.W. Gibbs

Humans effortlessly perceive the shape and motion of unfamiliar objects from their changing images. From a purely theoretical point of view this ability of the human visual system has intrigued perception psychologists for decades, and more recently computer vision scientists, for two reasons. First, valuable information is lost due to the projection of the three-dimensional scene onto the two-dimensional retina, and second, light as an intermediary of information transmission introduces certain ambiguities (e.g. the *aperture problem* discussed later) into the interpretation process. A primary goal of this research is to understand this phenomenon at the level of its *computational theory* (Marr, 1982). A computational study of this phenomenon is usually carried out in two stages. First is the *measurement of visual motion.* This involves computing the motion of image elements from the changing intensity pattern on the eye’s retina. The second stage is the *interpretation* of this visual motion, i.e., to infer the three-dimensional shape and motion of objects given the visual motion. The first stage, the measurement of visual motion, has been intensively studied by a number of researchers (e.g.: Horn and Schunck, 1981; Hildreth, 1983; Waxman and Wohn, 1985). This book is concerned with the second stage, the interpretation of visual motion. It extends the previous work of Longuet-Higgins and Prazdny (1980) and Waxman and Ullman (1985) in this area in many ways. A general formulation of the problem and algorithms for the interpretation process are presented.

An important goal of computer vision is to understand human vision from an information processing perspective. For this purpose, the task of vision can be divided into several stages, at least as a first approximation. Marr has suggested that the goal of the first stage of visual processing is to obtain descriptions of the physical properties of visible surfaces with respect to the viewer, properties such as distance, orientation, texture, and reflectance. This stage has been termed the *2 1/2 -D sketch* and the processes involved are called *early vision processes*. This early stage of processing is primarily bottom-up, relying on general knowledge about the world, but not on special high-level information about the scene to be analyzed. Computational studies and perceptual experiments (Marr and Poggio, 1977) suggest that early vision processes are generic ones that correspond to conceptually independent modules that can be studied, at least to a first order, in isolation. Examples of early vision processes are *edge detection* for finding sharp intensity changes, *stereopsis* for computing a depth-map from a stereo pair of images, *shape from shading*, *shape from texture*, *measurement of visual motion* and *interpretation of visual motion*.

There is no proof yet that the paradigm for computational vision proposed by Marr and his collaborators is correct, but we adopt it in the belief that something similar should be true. In this framework we contend that a rigorous and thorough analysis of the individual visual modules is fundamental to understanding vision as an information processing task. For this reason, this study focuses on one module, the visual motion module. Existence of this module as an independent process in the human visual system is demonstrated in many perceptual studies (Wallach and O'Connell, 1953; Johansson, 1973, 1975; Ullman, 1979; see Figure 1).

Another very important motivation for this research arises from its potential applications in machine vision systems. This work is directly relevant to autonomous land vehicle and aircraft navigation, robot manipulation of moving machine parts, and general machine vision systems.

This thesis is a *computational study* of the problem of visual motion interpretation. Before proceeding further we define the two key terms *computational theory* and *computational approach*. These two terms were originally elucidated by Marr.

A formulation of a *computational theory* of an information processing task consists of four steps:

(i) identifying the input and the output entities,

(ii) specifying the relationship between the input and the output entities,

(iii) explicitly stating the conditions and assumptions under which the output entities are obtainable from the input entities using the relations specified in step two, and

(iv) proving that the output entities are indeed obtainable from the input entities under the conditions and assumptions stated in step three using the relations specified in step two.

A *computational approach* to an information processing task consists of two steps:

(i) constructing a representation for the input and the output entities and

(ii) developing an algorithm to transform input into output according to a computational theory of the task for the representation constructed in step one.

In addition to the above two steps, we would like to add a third requirement for any computational approach:

(iii) an implementation of steps one and two on some processing hardware, such as a computer system, and a demonstration of its correctness on a variety of cases.

Tsotsos (1987) has recently argued that a computational study of a task must also include a *complexity level analysis.*

For the problem of visual motion interpretation, the input is the visual motion field obtained from changing intensity patterns on the eye's retina and the output is a description of the changing scene. In this thesis a computational theory of this process is formulated and a computational approach is given in the precise sense defined above. In this approach the input representation is a set of *image parameters* and the output representation is a set of *scene parameters*. The image parameters describe the local visual motion in the image domain; these are the ‘‘observables’’ or ‘‘knowns’’. The scene parameters describe the local shape and motion of surfaces in the scene; these are the ‘‘unknowns’’. Therefore the goal of visual motion interpretation is to recover scene parameters from image parameters. Relations between the scene parameters and the image motion parameters are established and an algorithm is given to recover scene parameters from image motion parameters. The algorithm is implemented and tested on a computer system.

The process of inferring the time-varying geometry of a scene from the corresponding visual motion is carried out locally, both in space and in time. The reason for this is that it is impractical to represent arbitrary shapes and motions of surfaces in the scene by a global parameterization scheme. Restriction to local analysis facilitates working in a smaller parameter space (but with less information, of course). Even this local analysis is assumed to be preceded by a detection of discontinuities in the visual motion corresponding to discontinuities in the geometry (distance, orientation, curvature, etc) and the motion (translation, rotation, acceleration, etc.) of surfaces in the scene. This can, in principle, be achieved by an appropriate modeling of the visual motion (e.g. requiring the motion field to be described by polynomials of fixed degree up to a preset tolerance). Having located such discontinuities in the motion field a local analysis is carried out in small image regions not containing these discontinuities to recover the structure and motion of the corresponding surfaces in the scene. A patching together of this local three-dimensional information is necessary to obtain a global description of the scene.

The perception of motion from a monocular visual stimulus is investigated. A computational theory of the interpretation of visual motion caused by the projection of a moving surface is presented. The formulation of the theory is basically an extension of the earlier formulations of Longuet-Higgins and Prazdny (1980) and Waxman and Ullman (1985). Further, a computational approach, in the sense defined earlier, is given for the interpretation process.

A major portion of this book is devoted to the study of visual motion resulting from rigid motion of objects. Here the problem is to determine the three-dimensional shape and rigid motion of surfaces from their image motion. Equations relating the local surface parameters (slopes and curvatures) and motion parameters (translation and rotation) to the spatial image motion parameters are derived. These equations are solved for planar surfaces and curved surfaces; in both cases the solution is derived in *closed form.* The solutions show that, in general, planar surfaces have two interpretations whereas for curved surfaces the interpretation is unique. Many theorems concerning the multiplicity of interpretations are proved.

This formulation for the analysis of instantaneous image motion is then extended to the analysis of spatio-temporal image motion. In this formulation the equations relating the local orientation and motion of a surface and the first order spatio-temporal image motion derivatives are derived. For this case, again, the solution for the orientation and motion is derived in closed form. Further, an interesting case where a camera tracks a point on a moving surface is solved with the knowledge of the camera's tracking motion. An extension of this formulation to deal with non-uniform or accelerated motion is described. This extension is illustrated with a simple example.

Finally the formulation for rigid motion is generalized to deal with non-rigid motion. This again is illustrated with a simple example. This general formulation leads to some new insights into the intrinsic nature of the image motion interpretation problem. It makes explicit the well known fact that the general problem of inverting the perspective projection transformation is *inherently ill-posed* (or under-constrained), and that additional assumptions about the physical world are necessary to solve the problem. It gives the minimum number of additional constraints (in the form of assumptions about the scene) necessary to solve the problem. For example, it exposes the fact that the *rigidity assumption*, the assumption that objects in the scene are rigid, is a powerful and sufficient constraint that results in a unique interpretation in most cases. The general formulation serves to address the two fundamental issues: *What information is contained in the image motion field? How can it be extracted?*

We emphasize here that the approach taken in this thesis is based on the assumption that the physical phenomena affecting the visual motion, phenomena such as surface structure, translation, rotation, etc., vary ‘‘smoothly’’, both in space and in time. However this assumption is not a drawback as long as locations of discontinuities in the image motion field can be detected.

The computational approach developed here has been implemented on a computer system and tested on a variety of cases to verify the approach. Many experimental results are included.

An important conclusion of this thesis is that an independent perceptual module for visual motion processing is capable of providing a wealth of information under some natural assumptions. The study also discloses the limitations of such a module, limitations in the form of assumptions about the world that are necessary to extract scene information from visual motion.

The computational theory of visual motion perception presented here provides a basis for further investigations. It motivates an inquiry into the *second order theory* of the visual motion processing module, i.e. the interaction of this module with other visual processing modules. It naturally raises certain questions: What types of interactions are necessary? What type of interactions are possible? How can the different modules cooperate to achieve their goals efficiently and robustly?

An important outcome of this work is a computational approach which is potentially useful in machine vision systems. It provides a uniform representation and a unified algorithm applicable in a variety of situations. The approach is flexible in that *a priori* information can be easily incorporated in the form of additional constraints and it is general enough to be extensible to many situations not considered explicitly in this study. However, in order to appply the theory developed here to practical applications, visual motion needs to be measured very accurately. Accurate and robust measurement of visual motion has remained a difficult problem to this day.

An overview of the computational stages in visual motion analysis is given in the next chapter. It defines some essential terminology and summarizes previous work in this area. Chapter 3 gives the formulation of the problem for rigid motion of surfaces. Image motion equations that relate the scene parameters and the image parameters are derived. Chapter 4 deals with solving the image motion equations and the multiplicity of interpretations. Chapter 5 extends the analysis to the temporal domain. Spatio-temporal parameters of image motion are related to the three-dimensional structure and motion parameters. A number of interesting cases are solved including a simple case of non-uniform or accelerated motion. Chapter 6 generalizes the formulation to the non-rigid motion case. Error sensitivity analysis of the computational approach is discussed in Chapter 7.

APPENDIX A

Expressing a surface in terms of image coordinates

Here we give a method of deriving the function that maps image points (x,y) on the image plane to points on the surface in the scene along the optical axis. Since the image at a point (x, y) on the image plane corresponds to the point (xZ, yZ, Z) in the scene, our goal is to express Z in terms of (x, y) and the surface structure parameters. In Longuet-Higgins and Prazdny (1980) and Waxman and Ullman (1985) Z was so expressed only up to second order terms of (x, y). Below we give a systematic method which can be used to express Z up to any desired order of terms in (x, y).

Assuming that the surface is smooth and is given by $Z = f\,(X,Y)$ we can expand $f\,(X,Y)$ in a Taylor series:

$$Z = a_0+a_1X+a_2Y+a_3X^2+a_4XY+a_5Y^2+a_6X^3+..... \qquad \text{(A1)}$$

Using equations (3.1a,b) and equation (A1) we can obtain an implicit expression for Z in terms of the image coordinates x, y :

$$Z = a_0 + Z\,(a_1x + a_2y + \underline{Z}\,(\,a_3x^2 + a_4xy + a_5y^2 + Z\,(a_6x^3 + ...)\,)\,) \qquad \text{(A2)}$$

Now we systematically substitute for the appropriate Zs on the right hand side to eliminate second and higher order terms in Z on the right hand side of equation (A2). Substituting the entire right hand side of equation (A2) for the Z underlined in equation (A2) we get

$$Z = a_0+Z\;(a_1x+a_2y+(a_0+Z\;(a_1x+a_2y+....)) \qquad \text{(A3)}$$

$$(a_3x^2+a_4xy+a_5y^2+Z\;(a_6x^3+.....)))$$

Rearranging terms in equation (A3) we have

$$Z = a_0+Z(a_1x+a_2y+a_0a_3x^2+a_0a_4xy+a_0a_5y^2+a_0\underline{Z}(a_6x^3+.....) \qquad \text{(A4)}$$

$$+\underline{Z}(a_1x+a_2y+....)(a_3x^2+a_4xy+a_5y^2+Z(a_6x^3+.....)))$$

We again substitute for the Zs underlined in equation (A4) the entire expression on the right hand side of equation (A4):

$$Z = a_0 + Z(a_1 x + a_2 y + a_0 a_3 x^2 + a_0 a_4 xy + a_0 a_5 y^2 + (a_0^2 a_6 + a_0 a_1 a_3) x^3 + \ldots). \qquad \text{(A5)}$$

Continuing this recursive substitution procedure, Z can be expressed explicitly in terms of the image coordinates x, y to any required order of terms. Using $O_3(x,y)$ to denote third and higher order terms equation (A5) can be written as

$$Z = a_0 + Z\ (a_1 x + a_2 y + a_0 a_3 x^2 + a_0 a_4 xy + a_0 a_5 y^2 + O_3(x,\ y))\ . \qquad \text{(A6)}$$

Rearranging terms in equation (A6) we get

$$a_0 = Z\ (1 - a_1 x - a_2 y - a_0 a_3 x^2 - a_0 a_4 xy - a_0 a_5 y^2 - O_3(x,y))\ . \qquad \text{(A7)}$$

Equation (A7) can be used to obtain an expression for Z which is explicit up to second order terms:

$$Z = a_0\ [1 - a_1 x - a_2 y - a_0 a_3 x^2 - a_0 a_4 xy - a_0 a_5 y^2 - O_3(x,y)]^{-1}\ . \qquad \text{(A8)}$$

Contents