THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE

ROBOTICS: VISION, MANIPULATION AND SENSORS

Consulting Editor

Takeo Kanade

Other books in the series:

ROBOTIC GRASPING AND FINE MANIPULATION, M. Cutkosky
ISBN: 0-89838-200-9

SHADOWS AND SILHOUETTES IN COMPUTER VISION, S. Shafer
ISBN: 0-89838-167-3

PERCEPTUAL ORGANIZATION AND VISUAL RECOGNITION, D. Lowe
ISBN: 0-89838-172-X

ROBOT DYNAMICS ALGORITHMS, F. Featherstone
ISBN: 0-89838-230-0

THREE- DIMENSIONAL MACHINE VISION, T. Kanade (editor)
ISBN: 0-89838-188-6

KINEMATIC MODELING, IDENTIFICATION AND CONTROL OF ROBOT MANIPULATORS, H.W. Stone
ISBN: 0-89838-237-8

OBJECT RECOGNITION USING VISION AND TOUCH, P. Allen
ISBN: 0-89838-245-9

INTEGRATION, COORDINATION AND CONTROL OF MULTI-SENSOR ROBOT SYSTEMS, H.F. Durrant-Whyte
ISBN: 0-89838-247-5

MOTION UNDERSTANDING: Robot and Human Vision, W.N. Martin and J. K. Aggrawal (editors)
ISBN: 0-89838-258-0

BAYESIAN MODELING OF UNCERTAINTY IN LOW-LEVEL VISION, R. Szeliski
ISBN 0-7923-9039-3

VISION AND NAVIGATION: THE CMU NAVLAB, C. Thorpe (editor)
ISBN 0-7923-9068-7

TASK-DIRECTED SENSOR FUSION AND PLANNING: A Computational Approach, G. D. Hager
ISBN: 0-7923-9108-X

COMPUTER ANALYSIS OF VISUAL TEXTURES, F. Tomita and S. Tsuji
ISBN: 0-7923-9114-4

DATA FUSION FOR SENSORY INFORMATION PROCESSING SYSTEMS

by

James J. Clark
Alan L. Yuille

Division of Applied Sciences
Harvard University
Cambridge, Massachusetts

KLUWER ACADEMIC PUBLISHERS
Boston/Dordrecht/London

Distributors for North America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA

Distributors for all others countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS

Library of Congress Cataloging-in-Publication Data

Clark, James Joseph, 1957–
Data fusion for sensory information processing systems / James J. Clark, Alan L. Yuille.
p. cm. — Kluwer international series in engineering and computer science ; SECS 105)
Includes bibliographical references (p.) and index.
ISBN 0-7923-9120-9 :
1. Computer vision. 2. Image processing. I. Yuille, A. L. (Alan L.) II. Title. III. Series.
TA1632.C58 1990
006.3 '7—dc20 90-4770
CIP

Printed in the United States of America

Contents

Preface xiii

1 Introduction: The Role of Data Fusion in Sensory Systems 1

1.1 INFORMATION ACQUISITION: INVERTING THE WORLD-IMAGE MAPPING 1

1.2 THE NEED FOR CONSTRAINTS 5

1.3 DETERMINATION AND EMBEDDING OF CONSTRAINTS 9

1.4 THE NEED FOR DATA FUSION 13

1.5 SUMMARY 15

2 Bayesian Sensory Information Processing 17

2.1 BAYES RULE 18

2.2 THE IMAGE FORMATION MODEL 19

2.3 THE PRIORS 23

2.3.1 THE SYSTEM MODEL 24

2.4 BAYESIAN ESTIMATORS FOR $\vec{f}$ 25

2.5 BAYESIAN DETECTION AND EXTRACTION SYSTEMS 28

2.6 THE BAYESIAN CONTROVERSY 30

2.7 CHAPTER SUMMARY 37

3 Information Processing Using Energy Function Minimization 39

3.1 MARKOV RANDOM FIELDS 44

3.2 ENERGY FUNCTIONS WITH MATCHING ELEMENTS 46

3.3 STATISTICAL MECHANICS AND MEAN FIELD THEORY 50

3.3.1 W.T.A. WITH THE MEAN FIELD APPROXIMATION 51

3.3.2 W.T.A. WITHOUT THE MEAN FIELD APPROXIMATION 53

3.3.3 AVERAGING OUT FIELDS 54

3.4 THE FORM OF THE SMOOTHNESS CONSTRAINT 55

3.5 ALTERNATIVE FORMS OF CONSTRAINT 57

3.5.1 PARAMETRIC CONSTRAINTS 57

3.5.2 MINIMAL DESCRIPTION LENGTH CODING 64

3.5.3 MULTIPLE SETS OF PRIORS 66

3.6 CHAPTER SUMMARY 68

4 Weakly vs. Strongly Coupled Data Fusion: A Classification of Fusional Methods 71

4.1 A CLASSIFICATION OF FUSIONAL METHODS . . 72

4.2 WEAKLY COUPLED DATA FUSION 73

4.2.1 CLASS I WEAKLY COUPLED DATA FUSION 73

4.2.2 CLASS II WEAKLY COUPLED DATA FUSION 75

4.2.3 CLASS III WEAKLY COUPLED DATA FUSION 76

4.3 STRONGLY COUPLED DATA FUSION ALGORITHMS 78

4.3.1 STRONG COUPLING BY PRIOR CONSTRAINT ADAPTION 80

4.3.2 STRONG COUPLING BY ADAPTION OF THE IMAGE FORMATION MODEL 80

4.3.3 RECURRENT STRONG COUPLING 81

4.3.4 COUPLED MRF METHODS AS STRONGLY COUPLED DATA FUSION 82

4.4 BAYESIAN IMPLEMENTATION OF DATA FUSION 83

4.5 EXAMPLES OF WEAKLY COUPLED DATA FUSION IN THE VISION LITERATURE 87

4.6 EXAMPLES OF STRONGLY COUPLED FUSION IN THE VISION LITERATURE 91

4.7 SUMMARY 103

5 Data Fusion Applied to Feature Based Stereo Algorithms 105

5.1 INTRODUCTION 105

5.2 THE BAYESIAN APPROACH TO STEREO VISION 108

5.2.1 THE MATCHING PROBLEM 108

5.2.2 THE FIRST LEVEL: MATCHING FIELD AND DISPARITY FIELD 109

5.2.3 THE SECOND LEVEL: ADDING DISCONTINUITY FIELDS 112

5.2.4 THE THIRD LEVEL: ADDING INTENSITY TERMS 113

5.2.5 THE BAYESIAN FORMULATION OF THE STEREO ALGORITHM 114

5.3 STATISTICAL MECHANICS AND MEAN FIELD THEORY 116

5.3.1 AVERAGING OUT FIELDS 117

5.3.2 DETERMINISTIC SOLUTIONS OF THE MEAN FIELD EQUATIONS 125

5.4 COMPARISONS WITH OTHER THEORIES 128

5.4.1 THE MARR-POGGIO COOPERATIVE STEREO ALGORITHM 128

5.4.2 DISPARITY GRADIENT LIMIT THEORIES 130

5.5 COMPARISONS WITH PSYCHOPHYSICAL DATA 131

5.6 CHAPTER SUMMARY 134

6 Fusing Binocular and Monocular Depth Cues 137

6.1 STRONG FUSION - STEREO WITH MONOCULAR CUES 137

6.2 PREVIOUS ATTEMPTS AT STRONG COUPLING FOR STEREO 138

6.2.1 A GENERAL FRAMEWORK 142

6.2.2 SOFT AND HARD CONSTRAINTS 144

6.3 SUMMARY 146

7 Data Fusion in Shape From Shading Algorithms 147

7.1 AN ALGEBRAIC APPROACH TO FUSING SPECULAR AND LAMBERTIAN REFLECTANCE DATA 148

7.2 A CLASS III WEAKLY COUPLED FUSION IMPLEMENTATION 158

7.3 A STRONGLY COUPLED APPROACH TO POLYCHROMATIC SHAPE FROM SHADING . . . 168

7.4 FUSION OF IMAGE FORMATION MODELS 172

7.5 CHAPTER SUMMARY 179

8 Temporal Aspects of Data Fusion 181

8.1 A TEMPORAL COHERENCE EDGE DETECTOR . 182

8.1.1 DETERMINATION OF THE CONDITIONAL DENSITIES 185

8.1.2 BAYESIAN EDGE DETECTION DECISION PROCESS 189

8.2 A STRONGLY COUPLED TEMPORAL COHERENCE EDGE DETECTOR 197

8.3 TEMPORAL SAMPLING 201

8.3.1 COMPUTATIONAL CONSTRAINTS 204

8.4 ACTIVE DETERMINATION OF CONSTRAINTS . 206

8.5 SUMMARY 215

9 Towards a Constraint Based Theory of Sensory Data Fusion 217

Bibliography 223

Index 239

Preface

The science associated with the development of artificial sensory systems is occupied primarily with determining how information about the world can be extracted from sensory data. For example, computational vision is, for the most part, concerned with the development of algorithms for distilling information about the world (e.g. localization and recognition of various objects in the environment) from visual images (e.g. photographs or video frames). There are often a multitude of ways in which a specific piece of information about the world can be obtained from sensory data. A subarea of research into sensory systems has arisen which is concerned with methods for combining these various information sources. This field is known as *data fusion,* or *sensor fusion.* The literature on data fusion is extensive, indicating the intense interest in this topic, but is quite chaotic. There are no accepted approaches, save for a few special cases, and many of the best methods are *ad hoc.*

This book represents our attempt at providing a mathematical foundation upon which data fusion algorithms can be constructed and analyzed. The methodology that we present in this text is motivated by a strong belief in the importance of constraints in sensory information processing systems. In our view, data fusion is best understood as the embedding of multiple constraints on the solution to a sensory information processing problem into the solution process. Different data fusion algorithms, if one takes the point of view espoused in the text, can be differentiated by the ways in which constraints are embedded, and by the reasons why the constraints are necessary.

The constraint based approach that we take to the study of data fusion leads to a classification of data fusion algorithms into two general types: weakly coupled and strong coupled. Members of these classes are distinguished by the manner in which information, in the form of constraints on the solution, are combined in order to obtain a solution to a sensory information processing problem. In weakly coupled data fusion algorithms the data from a set of sensory

processing modules is combined in a way that does not affect the operation of any of the modules. In strongly coupled algorithms the outputs of sensory processing modules are allowed to interact with other modules and affect their progress. Typically this alteration of modules by other modules is done by altering the constraints or assumptions used by a given sensory module (or source of sensory information) based on the output of some other module or sets of modules.

The bulk of the theoretical development and examples of the application of the theory presented in this book assume a Bayesian formulation of sensory processing modules. This approach is taken since the ideas inherent in our classification of data fusion systems can be most clearly understood in terms of Bayesian methodology. Our theories, however, are by no means limited to sensor processing algorithms based on Bayesian methods.

Chapter 1 is a introduction to the constraint based view of sensory information processing systems. We describe the general classes that these constraints can be put into, and examine the problem of determination and embedding of the constraints into a system.

Chapter 2 provides a review of Bayesian information processing methods. The relationship of the Bayesian methodology to the generation and embedding of constraints is detailed along with a brief historical summary of the scientific and philosophical controversy surrounding it.

In Chapter 3 we introduce the energy function minimization approaches to information processing and the relationship of these methods to Bayesian techniques is outlined. Included in our discussion of energy function methods are the application of Markov random field theory and the description of information processing techniques derived from the application of mean field theory to the energy function minimization process.

Chapter 4 forms the heart of the text, as it describes our constraint based approach to data fusion. We describe the role of data fusion in reducing the role of natural and artificial constraints in sensory processing operations and detail how we can use both the Bayesian and the related energy function formalisms to provide natural implementations of data fusion. We introduce a fundamental characterization of data fusion algorithms as implementing either weakly coupled or strongly coupled methods. Weak coupling refers to fusion of data produced by sensory modules in ways that do not affect the operation of the modules, while strong coupling refers to fusional processes wherein the modules producing the data to be fused are being affected in some way by other information from other modules.

We detail the theory behind weakly coupled data fusion algorithms, and provide a utility based classification of these algorithms. The Bayesian interpretation/implementation of weakly coupled data fusion is described. The chapter continues with a number of examples, drawn from the vision literature, of weakly coupled fusional algorithms applied to computational vision. These examples are intended to illustrate the range of fusional techniques and how they relate to the classification given earlier in the chapter. This review of data fusion algorithms found in the computational vision literature shows the prevalence of weakly coupled data fusional methods in the field.

In Chapter 4 we also cover the theory of strongly coupled data fusion techniques. The different classes of strong coupling, namely - prior model adaption, image formation model adaption, and recurrent coupling, are examined in detail. The theory behind coupled Markov random field methods is revisited and the relationship between these approaches and strong coupling is shown. The relationship of mean field approaches to energy function based information processing and the Bayesian interpretation of strong coupling is detailed. Some examples of vision algorithms found in the vision literature that are effectively strongly coupled are described.

Chapters 5 through 7 include a number of examples of the use

of weakly and strongly coupled data fusion to important problems in computational vision. Chapter 5 illustrates the use of the energy function based methods developed in Chapter 3 in developing new algorithms for performing stereo vision. Chapter 6 describes the use of similar methods for constructing algorithms that can fuse binocular and monocular object shape information. Chapter 7 is an in depth study of various weakly and strongly coupled fusional methods for computing object shape from the Lambertian and specular components of the light reflected from the object.

In Chapter 8 we investigate some of the temporal aspects of data fusion. These include the rectification, or resampling, of data in the time domain. An important issue in the design of fusional systems is the effect of temporal constraints, such as the processing latency and throughput of sensory modules, on the performance of fusional systems. We discuss a form of (strongly coupled) data fusion wherein image formation model parameters under the control of the algorithm are changed over time in a way that improves the performance of the fusional algorithm. The relationship of such "active fusion" techniques to what is known in the literature as "active vision" is examined. The chapter concludes with a detailed description of the use of an important constraint available in processing time varying image data, that of the temporal coherence of image features.

The final chapter summarizes the information presented in the book and suggests a constraint based theory of sensory data fusion. The theory is based on our conviction that constraints are of primary importance in the solution of sensory information processing tasks.

At the end of each chapter is a brief summary of the main ideas presented in the chapter. It is intended to permit the casual reader to see what is in the chapter before they dive in amongst the mathematics.

We would like to conclude our prefatory remarks by acknowledging the influences that a number of people have had on our work. The contributions of Ten-Lee Hwang in terms of ideas and comments, especially concerning temporal aspects of data fusion, were much appreciated by the authors. The mean field theoretic ideas proposed in this book were developed by the second author in collaboration with Davi Geiger, who was, in addition, a source of many ideas and diversions. Norberto Grzywacz was a goldmine of information regarding neurophysiological evidence for different approaches to visual processing modules. David Mumford's work on the energy function minimization approach to vision was a source of inspiration, as were the many discussions the authors had with him on related topics. Ken Keeler provided a useful sounding board for ideas concerning the application of minimal length encoding principles to data fusion. Roger Brockett rolled his eyes heavenward when he heard we were wasting our time writing this book, but was nonetheless an enthusiastic supporter of the program of research detailed herein. The first author would like to thank Alan Mackworth for introducing him to the importance of constraints in computational vision.

Conversations (and implementational work done with) with the following people about computational vision contributed greatly to the development of the ideas in this text: Heinrich Bülthoff, David Cohen, John Daugman, Nicola Ferrier, Mike Gennert, Gaile Gordon, Peter Hallinan, Tai-Sing Lee, Pamela Lipson, Petros Maragos, Charlie Marcus, Mark Nitzberg, Ed Rak, Taka Shiota, John Wyatt, and Tong Yang. Bill Nowlin, besides writing the great software with which the graphs in the text were printed, made known to us various points regarding products of probability densities. We would like to thank Rick Szeliski for providing us with his LaTeX index making utilities left over from the writing of *his* book, and, more importantly, for discussions about data fusion and energy minimization techniques. Finally, the authors would like to thank George Thomas for keeping our computers running smoothly and not allowing the disks to crash.

The writing of this book and the research described in it was performed with the support of the Brown-Harvard-MIT Center for Intelligent Control Systems, funded by the Army Research Office, grant number DAA103-86-K-0171. Beyond the generous financial

support, the Center provided fruitful interactions with the vision groups at Brown University and the Massachusetts Institute of Technology. Our research was influenced in no small part by the work of Stuart Geman, Basilas Gidas and Donald McClure of Brown University and by the work of Sanjoy Mitter and colleagues at MIT. We would like to thank Jagdish Chandra of the Army Research Office for his support of the Center.

Significant support was also provided through the Maryland-Harvard NSF Center of Excellence in Systems Research, grant number CDR-85-00108.

James J. Clark and Alan L. Yuille
Cambridge, Massachusetts
March 1990

DATA FUSION FOR SENSORY INFORMATION PROCESSING SYSTEMS

Chapter 1

Introduction: The Role of Data Fusion in Sensory Systems

1.1 INFORMATION ACQUISITION: INVERTING THE WORLD-IMAGE MAPPING

Living organisms are distinguished by their information gathering abilities. Much effort has been expended by biologists and other scientists in determining by what means organisms obtain and process information. Recently, scientists interested in the computational aspects of information processing have looked towards these studies in search of clues as to how machines can be made to gather and process information. In this book we focus on an important aspect of sensory information processing, that of fusing separate sources of sensory information.

We begin our study by asking the fundamental question of why

organisms need to gather information. If we take the controversial view of J.J. Gibson regarding affordances [54], we can answer the above question quite simply. In this view organisms need information about their environment in order to carry out actions particular to the goals of the organism. The information required is perception, or recognition and localization, of structures in the environment of the organism which *afford* the activities that the organism wishes to carry out. These affordances are obtained by examining sensory information involving the spatial-temporal form of and relationships between world structures.

The affordance view may be quite valid for simple systems or organisms, and is certainly a useful concept for robotic systems (which are, by their nature, motivated by concerns of utility), but falls short in explaining the reasons for information gathering in complex organisms such as mammals. In higher organisms, information is not always related to the utility of objects with respect to the goals of the organism. For example, distinctions may be needed to be made between structures of equivalent utility (e.g. different styles of chairs). This requires some knowledge about the shape or form of structures in the world. There is evidence that in humans there are separate systems for dealing with utility of world structures and with the shape of these structures [155] (see the interesting discussion on this point in Marr's book on vision, [97], p. 35). Higher level concepts such as quality and beauty of world structures may also be required to be extracted from the sense data, although such high level cognitive concepts can conceivably be interpreted in terms of the affordance model (e.g. a beautiful object affords enjoyment or appreciation).

In either case, it is clear that organisms need to extract information of some kind about structures in their environment. This information is either innate (or hard-wired) or must come from sense data. In order to understand the information gathering process we must understand how innate knowledge is "learned" through studies of the evolutionary process, and we must understand how sense data is generated from arbitrary configurations of physical structures in the environment. We will leave the first task to the biologists and

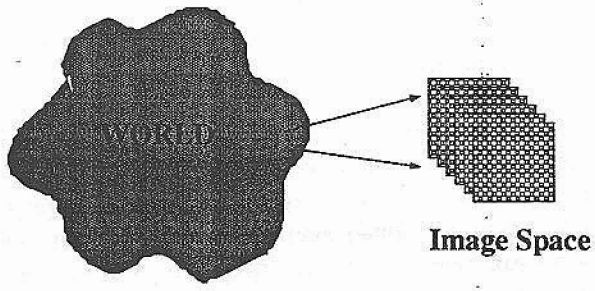


Figure 1.1: The world-image mapping

will concentrate on the second (although we definitely will not ignore innate or learned knowledge sources in our discussions).

Let us define the *sensing* process as that by which measurements of quantities (sense data), which are dependent on the structures in the world or environment and their configuration, are obtained. The *perception* process can be defined as the process of deriving from the sense data specific items of information about meaningful structures in the environment. The sensing process can be interpreted as a mapping of the state of the world into a set of images of much lower dimensionality. We will call this mapping the *world-image mapping*. This mapping is depicted in figure 1.1. The usual operations involved in the world-image mapping are sampling and projection. Both of these processes result in loss of information. One way of stating this is that the mapping from the space of world structures to the image space is many-to-one. This means that there are typically many possible (usually an infinite number) configurations of the world that could give rise to the measured sense data. A simple example of this is the well known Necker Cube, shown in figure 1.2. Most observers of this image would agree that there are at least two interpretations of the world structure that gave rise to it. These are the interpretation of the image as being the image of a three dimensional wire model of

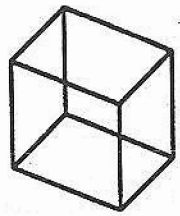


Figure 1.2: A simple example of a non-invertible world-image mapping: the Necker Cube.

a cube, with the difference between the two interpretations being due to the ambiguity about which corner is in front of which. More astute observers would grant the existence of a third possibility, that of the object being flat or two dimensional. If one thinks about the situation a little more deeply they should realize that there are actually an infinite number of possible interpretations of the image. For example, the object could be a wire frame outline of a non-cubical object which has been rotated so as to appear like a cube. There are an infinite number of such rotations. There are also aspects of the world which may not be accessible through the available sense data. For example, the wire frame object may be made out of an infinite variety of different materials, all of which appear the same. The object could be occluding other objects, which form part of the world but cannot be sensed. Even worse, the image shown in figure 1.2 is obviously a sampling of the world since it has a finite extent. We cannot say anything about the parts of the world that are not sensed. For example, the world configuration (or *scene)* that gave rise to figure 1.2 could be a wire frame representation of a cube (made of graphite) moving upwards to the right at 2 meters/second, three meters above a potted plant... etc. The inherent non-invertibility of the world-image map, even for cases which seem fairly straightforward, should

be quite evident.

We can treat the perception process as the process of inverting the world-image map. The world-image map is, as we have seen, non-invertible. Herein lies the primary difficulty which must be overcome in designing sensory information processing systems. How *do* we invert the world-image mapping? The answer to this lies in *constraining* the set of possible world configurations to the point where the mapping from this reduced space to the image space *is* invertible. This idea, which forms the basis of modern computational vision theory, is examined in greater detail in the next section.

1.2 THE NEED FOR CONSTRAINTS

One of the most important findings of research into sensory systems, especially computational vision research, has been that most, if not all, simple sensory processing modules are inadequate for operation in anything more than a very constrained environment. They are inadequate in the sense that they often give the wrong answer as compared with the answer produced by the human visual system (which itself may be "wrong" as well). The primary reason for this is that *all* sensory information processing algorithms are necessarily based on constraints of one form or another that are imposed on the set of solutions to the sensory task being addressed by the module. A typical example is the surface smoothness constraint assumed in Grimson's [57] implementation of the Marr-Poggio [99] depth from binocular stereo algorithm. In this algorithm the assumption that surfaces are generally smooth (i.e. do not contain any discontinuities in depth) is used to choose between competing candidate matches between features in the two images. Other commonly used constraints in computer vision systems are the rigidity constraint used in motion analysis [150], the Lambertian assumption used in some shape from shading algorithms [62] and the polyhedral object constraint of blocks world object recognition systems [134]. If the environment that the sensory system is working in is sufficiently ge-

neral so that the constraints inherent in a given sensory processing module are violated some of the time, then that module will fail at these times, perhaps catastrophically.

The question arises, then, that if these constraints that are added to the vision algorithms are so troublesome, why are they needed at all? Is it not possible to design vision algorithms, or any sensory information processing system, that do not require any constraints or assumptions in order to function? This question is actually a fundamental one and cuts to the heart of not only the problem of developing sensory information processing systems but also that of science in general. As seen in the previous section, sensing involves a noninvertible projection of the state of the world. The sensory information processing task (or perception) is thus fundamentally that of solving an ill-posed problem and must be approached with the mathematical tools that have been developed to handle such inverse problems.

The only way in which we can solve the perception problem is to somehow make the world-image mapping invertible. This can be done, as we remarked earlier, by imposing enough constraints on the space of possible world configurations so that the mapping between this constrained space and the image space *is* invertible. We can use this approach in solving the perception problem. That is, we add constraints to the solution process that allow a unique solution to be obtained. Thus, computer vision researchers developing vision algorithms have supplemented the sense data with extra information in the form of (often *ad hoc) a priori* constraints. These constraints are not based on current sensory data. However, there is no guarantee that these constraints will be valid when they are applied to the vision problem at a given time. In fact, a large number of the constraints commonly used in computational vision modules are invalid at least some of the time even for highly constrained domains. These include the aforementioned surface smoothness constraint, object rigidity constraint and polyhedral object constraint.

There are, however, a number of constraints which are almost

always valid, even in the complex sensory environments that most organisms function in. These are constraints based on the laws of physics and mathematics, such as Euclidean geometry, Fermat's principle of least action, Newton's laws of motions, the laws of thermodynamics and so on. These constraints allow us to rule out possible solutions to the vision problem that are "physically impossible". It is an open philosophical and scientific question as to whether these physical constraints are universally valid; however, for the purposes of this text we will assume that there exist a set of "physical" and "mathematical" constraints which are universally valid. Without this assumption we can not even attempt to solve the vision problem. These physical constraints allow us to set up the vision problem but are not usually enough by themselves to permit the attainment of unique solutions to the vision problem. To do this we need to add additional "natural" and "artificial" constraints.

Natural constraints are constraints which are derived from observations of our environment and represent conditions that are "usually" true in a given domain. These constraints can be thought of as the products of a form of scientific inquiry, wherein hypotheses are made and tested with experiment (see the discussion on active sensing later in the text), or wherein results of experiments are explained with theories regarding the state of the world and of the sensing process. The importance of natural constraints to computer vision was emphasized by Marr [97]. Examples of natural constraints are surface smoothness, object rigidity and Lambertian surface reflectance. One of the most commonly assumed natural constraint (but rarely stated explicitly) is the general viewpoint or genericity constraint. This assumption says that the environment does not present any non-generic or coincidental arrangements of the world structures in a manner so as to "fool" the sensory system.

Artificial constraints are a means by which high level knowledge is incorporated into a sensory process. These constraints are further removed from the sensing process than the "natural" constraints and embed expectations about the state of the world obtained through reasoning about previous estimates of the state of the world. An

example of an artificial constraint is provided by the "Ames room" demonstration familiar to most students of psychology. In this demonstration a room is constructed with a slanted ceiling so that one end of the room is shorter than the opposite end. An observer looking through a window located in the tall side of the room at two people of equal height, one standing by the far (short) wall, the other standing in the center of the room, typically perceives the person near the far wall to be taller than the other person. This misperception is presumably due to the use of an (here invalid) artificial constraint that the ceilings of rooms are the same height everywhere in the room.

Some researchers have questioned the need for natural and artificial constraints in sensory information processing. Such an approach, championed by Gibson and his followers, arises from the "ecological optics" [54] view of vision in animals. In this approach artificial constraints are not needed as the additional information needed to invert the world-image mapping is thought, by proponents of this view, to be available from the "sensory flow" that arises as the animal moves about its environment. It is an open question, however, that the additional information available in this manner is sufficient to invert the world-image map uniquely without the need for any artificial constraints. Adherents of the Gibsonian view (see, for example, [108]) claim that the sensory flow is sufficient, but provide no proof for this claim. Also, they need at least to assume a general viewpoint or genericity constraint, as elements in the enviroment could conspire to produce misleading sensory flow.

Recent developments in computational vision, known by some as "active vision" [2, 1] have produced rigorous results on the ecological optics approach to information processing. These results appear to indicate that the information in the sensory flow can reduce the dependence on artificial and natural constraints, but have not shown that they can be eliminated entirely (at the very least general viewpoint constraints are needed).

Given that we absolutely need constraints in order to solve the perception problem, it now remains to determine the proper constraints to use, and how to embed them into our algorithms.

1.3 DETERMINATION AND EMBEDDING OF CONSTRAINTS

We are faced with two principal problems when considering the need for constraints in sensory information processing. These are

- What information (in the form of constraints or assumptions) do we need to know, in order to obtain the required world information uniquely from a given set of sensor data. We will call this the *constraint determination* problem.
- How can we embed the constraints into the sensory information processing algorithms? This is essentially an algorithmic design problem, and typically cannot be divorced from the question of how to use the sensory data to obtain the desired world information. We will call this the *constraint embedding* problem.

The determination of the proper constraints to use in a particular sensory information extraction task is a scientific pursuit. This should therefore form the basis of research into sensory processing systems, rather than *ad hoc* construction of sensory processing algorithms.

Philosophers have long debated whether knowledge can be obtained directly from experience or whether *a priori* organizing principles are necessary. For example, Kant [77] treated space and time as synthetic *a priori* intuitions that serve to begin the construction of knowledge from sense impressions. The manner in which a mental image is generated "is an art concealed in the depths of the human soul, whose real modes of activity nature is hardly likely ever to allow us to discover, and to have open to our gaze." Thus our "knowledge"

is a construct using *a priori* principles and true reality, the "Ding am Sich", is forever unknowable.

This approach can be contrasted with the British empiricist philosophers, Hobbes, Locke and Hume, [89, 104] who believed there was a closer connection between sense impressions and ideas. However, the flaw in pure empiricism was revealed by Hume's [104] question, "How do we know the sun will rise tomorrow?" The fact that it has always risen in the past, at least as far as we interpret our sense data, is no guarantee that it will rise in the future. Thus nothing can ever be proven by pure empiricism. A scientist may propose a law of nature which might agree perfectly with a million experiments, but it is not logically necessary that it will work for the million and one'th. In our terminology, just because a constraint has been proven to be valid for all situations experienced to date it does not follow that it will be valid for the next situation to be encountered. Therefore we should be careful in distilling constraints from prior experience.

The philosophical arguments presented above suggest that sense impressions are not sufficient in themselves for describing the world and obtaining natural laws. They must be supplemented by some *a priori* (or "innate") assumptions. The question of where these "innate" constraints come from then naturally arises. There is no clear answer to this question and it is still not clear whether or not non-experiential or innate prior knowledge is actually necessary. We will, in this book, take the view that if one sufficiently constrains the domain in which world descriptions are sought, then one *can* obtain suitable constraints solely through reasoning on sense data (i.e. through the application of the scientific method). The restriction of the domain is necessary in order to eliminate the need for the specification of "innate" unknowable (through sense data) *a priori* constraints.

The apparent impossibility of empirically proving a law of nature to be true (or universally valid) has lead some scientists to believe that laws should be thought of as an economical language for describing data rather than as being fundamental truths. For example,

Einstein argued that a theorist was like a tailor designing a suit to fit the phenomena; the alternative theories of gravity by himself and Newton were alternative styles. Risannen's ideas on minimal length encoding (see chapter 3) are reminiscent of these ideas of economical languages. Another way of stating this viewpoint is that any constraint (or piece of knowledge) is contingent and is only valid in a particular context; universally valid constraints do not exist (this argument is related to the argument concerning objective and subjective probabilities that is presented in chapter 2). In this paradigm, the affordance approach of Gibson is exactly the right one, as it says that the constraints to be used in a particular task are specified by the nature of the task. Thus we would conclude that the constraint determination problem requires a detailed analysis of the expected world or domain the system is to work in, *and* of the tasks that the system is intended to perform. The former analysis is aimed at deriving the natural and physical constraints associated with the chosen domain, while the latter is intended to weed out irrelevant constraints which are not associated with the required tasks.

This approach to sensory information processing is a *pragmatic* one, as we accept the fact that the constraints used are contingent on the particular domain and task set that we have chosen. If the domain is different or if it contains too many structures irrelevant to the tasks then we know that our perceptions may be incorrect, and we must hope that our system will not be too badly inconvenienced as a result. We assume that, for the most part, our assumptions will be valid and that our perceptions will be appropriate to the performing of the activities that are required.

Wherever the constraints needed to provide a unique solution to the problem of inverting the world-image mapping come from, one must have a way of embedding these constraints in the perceptual process. Computer vision researchers, whether they realize they are doing this or not, have come up with a number of ways of performing this embedding.

The most common way of embedding constraints into vision algorithms used during the early days of computational vision research was to simply restrict, or constrain, the domain in which the vision system is intended to work in. In effect, one is raising the level of a natural constraint to that of a physical constraint. One is playing God and imposing his own laws of physics on the domain. An illustrative example is that of the "blocks world" domain [58]. The designers of blocks world vision system decreed that non-polyhedral objects were physically impossible in their domain, and hence any vision system could safely assume that all objects it encountered would be polyhedral. A non-polyhedral object infringing into this domain would be as unnatural an occurrence as the presence of a pink elephant in our own environment. A large number of vision algorithms act in this manner. The problem with these early approaches to vision, however, was that the restriction of the domain was excessive, so that the resulting techniques worked poorly in the complex environments that any useful system would encounter.

Many of the more recent vision algorithms embed constraints in a more explicit manner. In the specification of these algorithms one distinguishes between physical and natural constraints, and models them separately. The algorithms are based on these models. In later chapters we will refer to the world model incorporating the physical constraints as the *image formation model,* while the natural constraints will be incorporated into both the image formation model and what we will refer to as the *prior* model. There are many techniques for constructing algorithms that use these models, and we will not be able to describe all of them in this text. We will concentrate on a particular type of technique which uses probabilistic forms of the models. The principal advantage of these probabilistic techniques is that the means by which constraints are embedded are particularly clear and intuitive (but not necessarily computationally efficient).

1.4 THE NEED FOR DATA FUSION

We have seen that the problem of inverting the world-image mapping is typically dealt with by using general assumptions (or natural constraints) about the world which are statistically true. However, as noted, it is not possible to obtain universally true constraints, so that there will always be some situations where our constraints are invalid. For example, in human vision illusions can occur when the assumptions used by the visual system are wrong. Nevertheless the human visual system is rarely fooled by illusions since a mistake made by one sensory information processing module is usually corrected by another module. Thus we could argue that, in order to obtain a domain independent sensory system, it is necessary to combine, or fuse, the results obtained from a number of modules. In this way, the constraints used by individual modules can be, to some extent, factored out in the fused sensory system so that the fused sensory system does not depend, or only weakly depends, on the assumptions made by its component modules. The idea here is that data fusion can reduce the dependence of the solution of a sensory information problem on invalid *a priori* constraints. In terms of inverting the world-image mapping we are, through the application of data fusion, replacing some of the *a priori* natural constraints, required for inversion, with independent data. This data can either be raw sense data, in which case we are expanding the dimensionality of the "image" space, or a partial inversion of the world-image mapping based on raw sense data. In the latter case we are reducing the dimensionality of the "world" space by constraining it with the result of the partial world-image mapping. In either case, we need fewer *a priori* constraints than if we did not use data fusion.

It is possible that, even if the assumptions or constraints used by a sensory module are valid, the module can still fail to produce reliable results. Often the output of a module will be unstable with respect to small changes in its inputs, such as may be produced by noise in a sensing device, and often the module is not able to come up with a unique solution (due to a lack of sufficient constraints). If

this is the case the sensory process being performed by the module is said to be ill-posed. We have seen that the addition of constraint can solve the lack of uniqueness problem, but, in many cases, stability can be assured as well by adding in suitable constraints on the solution. These extra constraints can be in the form of additional independent sensory information. Thus, we see that data fusion can be used to make an ill-posed problem well posed (or regularized [149]) in that it has a unique solution and that this solution is stable with respect to small perturbations in the data.

Another reason for the inadequacy of many sensory modules is that their outputs are noisy, or uncertain, even if all of their implicit assumptions are valid. By fusing the outputs of a given sensory module with those of other, independent, sensory modules one can reduce the uncertainty of the resulting measurement. This reduction of uncertainty arises from the averaging out of the independent noise processes that act on the different modules. If there is any dependence of the noise processes between modules, then the amount of averaging, and hence of the uncertainty reduction, will necessarily be reduced.

Based on the above discussion it should be evident that sensory fusion operations are useful in that they can do the following:

- Make a measurement problem well-posed, in that the resulting sensory process is stable to small perturbations in its input and that the system can provide a unique solution.
- Reduce dependence of a sensory module on possibly invalid assumptions or ad-hoc *a priori* constraints.
- Reduce the uncertainty in the value of a measured parameter.
- Reduce the effect of measurement noise on a quantity derived from the data.

The usual motivation for performing data fusion is to reduce the uncertainty in the value of some world parameter estimated from

sense data. One of the primary contributions of this text is to focus attention on data fusion as a means for reducing dependence on *a priori* natural and artificial constraints, rather than as a means for reducing uncertainty or minimizing the effects of noise. The techniques described in the text are useful in situations where the sense data are not very noisy, but where there is some uncertainty in the world models.

1.5 SUMMARY

- Information acquisition is required by organisms in order to determine the state of the environment that they are operating in. What information is required is related to the activities that the organism undertakes.
- In order to obtain the required world information one must invert the world-image mapping. In general this mapping is non-invertible so that extra information must be added to constrain the space of world configurations sufficiently to allow a unique solution to be obtained.
- This extra information is provided in the form of physical, natural, and artificial constraints. Physical constraints are derived from studies of the physical and mathematical laws underlying the world. Physical constraints are theoretically universally valid (although it may be impossible to discern their true form). Natural constraints are contingent on the particular restricted domain the organism is expected to function in. These constraints are not guaranteed, nor expected, to be universally valid. Artificial constraints are a form of natural constraint wherein the expectations are at a higher cognitive level. Artificial constraints are even less likely than natural constraints to be valid in a given situation.
- The determination of suitable constraints involves a characterization or modeling of the world and of the world-image

mapping. This is typically a scientific process. Once the constraints have been determined one must embed them into a suitable algorithm. The next chapter describes one such constraint embedding method, based on the application of Bayes rule.

- Natural or artificial prior constraints may be invalid in certain situations. In these cases one would like to reduce the dependence of a sensory information processing algorithm on these constraints if possible. One way of doing this is to use information from independent sensory information processing modules as constraints on a given module. This data fusion, then, is a means for reducing dependence of possibly invalid *a priori* constraints.

Chapter 2

Bayesian Sensory Information Processing

The introductory chapter addressed the need for the application of constraints, both natural and artificial, to aid in the performance of sensory information processing tasks. We saw that a major problem concerning the use of constraints lies in determining how the constraints are to be embedded in the algorithm(s) that carry out the information processing tasks.

A popular, and intuitive, approach to the embedding of constraints is based on a Bayesian interpretation of sensory information processing algorithms. In this approach different possible solutions are assigned probabilities of being the true solution based on prior expectations, and on models of the sensing process. The prior expectations can be influenced by previous measurements, as in the case of active vision or ecological optics, and by the constraints that we impose on the system. For example, imposing a surface smoothness constraint means that we expect that smooth surfaces are more likely than rough surfaces. We will make this idea more concrete in chapter 3, where we will show that one can express another common way of embedding constraints, that of energy function minimization (some-

times referred to as regularization) in the Bayesian framework. The energy minimization methods embed constraints by assigning solutions that are consistent with the constraints a lower energy than those that are not. Finding the solution having the lowest energy usually results in a solution that satisfies the constraints as well. The Bayesian approach also includes, and generalizes, other approaches such as those based on Markov Random Fields and Minimal Description Length (MDL) codes. We will discuss these generalizations, and their application to sensory information processing in the next chapter.

2.1 BAYES RULE

The Bayesian approach provides an elegant way of formulating sensory information processing problems in terms of probability theory. Bayes rule [10] relates the conditional probabilities of events:

$$Prob(X|Y) = \frac{Prob(Y|X)Prob(X)}{Prob(Y)}$$

Here X could be a specific visual scene, an apple for example. Y would be an image of the scene. The probabilities can be interpreted as: (i) $Prob(X|Y) = Prob(\text{Apple given the Image})$, (ii) $Prob(Y|X) = Prob(\text{Image given the Apple})$, (iii) $Prob(X) = Prob(\text{Apple})$, and (iv) $Prob(Y) = Prob(\text{Image})$.

Now we can specify $Prob(Y|X)$ by providing a model of how an apple reflects light. $P(X)$ is the *a priori* probability of there being an apple in any visual scene. $P(Y)$ is the *a priori* probability of the particular image. It is a normalization factor which can be determined from $Prob(Y|X)$ and $P(X)$.

In this framework the visual system needs to make assumptions to specify the *a priori* probability ($P(X)$) and the model of reflectance ($P(X|Y)$). Once these probabilities are determined the system can use Bayes rule to determine the probability of there being an apple given the image.

In the general sensory information processing case we will have a set of measurements, or data, from our sensing elements. Let us denote this set by a vector $\vec{d}$. A particular sensory information processing task would typically involve the determination, from $\vec{d}$, of some information about the world. Let us represent this desired world information by another vector $\vec{f}$ (this is implicitly assuming that the world information can be discretized in this way, an assumption we must make due to the inability of our computers to represent continuous information). We could then implement the sensory task in a Bayesian fashion by trying to determine $\vec{f}$ from some statistic of the following conditional probability:

$$P(\vec{f}|\vec{d}) = \frac{P(\vec{d}|\vec{f})P(\vec{f})}{P(\vec{d})} \tag{2.1}$$

where $P(\vec{f}|\vec{d})$ is the conditional probability that $\vec{f}$ is the proper solution given that we measure the data $\vec{d}$, $P(\vec{d}|\vec{f})$ is the conditional probability that $\vec{d}$ is measured given the function $\vec{f}$ and $P(\vec{f})$ is the *a priori* probability of $\vec{f}$.

2.2 THE IMAGE FORMATION MODEL

The conditional probability $P(\vec{d}|\vec{f})$ corresponds to a model of how world events having parameters $\vec{f}$ give rise to sense data $\vec{d}$. We will call this conditional probability the *image formation model*, using "image" to refer to any collection of sensory data[1]. The image formation model embeds physical constraints regarding how a sensing

[1]This is often referred to in the literature as the *sensor* model. We use the term image formation model to emphasize that the process of forming the image needs to be modeled just as much as the process of sensing the image. One should therefore think of our image formation model as both an image formation model and a sensor model.

element is affected by the world about it. It may include domain specific constraints, such as knowledge of the reflectance function of the object in a shape from shading algorithm, and it typically captures the uncertainty, or noise, in the sensor measurement and in our domain specific models. The precise form of the image formation model conditional probability depends on how the sensor noise, and the errors in the domain specific information are modeled.

The following example illustrates the specification of the image formation model. Suppose we are trying to find the shape of an object from the pattern of light reflecting from the surface of the object as received (sampled) by a camera (i.e. shape from shading). Our data $\vec{d}$ is the set of gray levels measured by the camera. The world features $\vec{f}$ that we wish to find are a set of surface normal vectors on the visible portions of the object. What then, is the form of the image formation model? We must first try to understand how the world features $\vec{f}$ map to the (noise free) sense data, call it $\vec{d}_0$. This is primarily a scientific endeavor, as it is essentially the determination of the proper physical constraints for the problem. Much attention has been given to the task of finding the laws governing the reflection of light off of bodies, and we will not summarize it here (but see [66] for such a summary). One can encompass the physical laws regarding reflection of light from surfaces with in a "Reflectance Map", which is a mapping from surface normal vectors to light intensity levels (and which depends on geometric factors regarding the direction of view and illumination sources, and on the intensity of the illumination sources). Thus we can say that the light reflected from the surface of an object, call it I, is given by:

$$I(\vec{x}) = R(\hat{n}(\vec{x}); \vec{s}, V(\vec{x}))$$

where $(\vec{x})$ is the coordinate of the point on the surface in question with respect to some coordinate system X, $\{\vec{s}\}$ is a set of illumination sources with intensity $|\vec{s}|$ and direction $\vec{s}/|\vec{s}|$, and V is some geometric information relating the position and orientation of the sensing elements to the object. The (noise free) sense data (call it $\vec{d}_0$) will be related to the reflected light pattern through some (possibly nonlinear) spatio-temporal operation $L_\Omega()$, where Ω is the spatio-temporal

support of the operator, and a coordinate transformation T. Thus:

$$d_0(\vec{y}) = L_\Omega[R(\hat{n}(\vec{x}); \vec{s}, V(\vec{x}))]_{T x \to y}$$

In practice one typically makes simplifying assumptions in modeling the surface to image mapping. For example, we assume a finite number of point light sources, all located at infinity, orthographic projection (essentially meaning that $V(\vec{x})$ is constant), Cartesian coordinate systems, flat sensing surfaces, point sensing elements with zero integration time (meaning that Ω is a point in space time), no mutual or self-illumination, and linear sensing elements (i.e. $L()$ is a linear operator) are all commonly used assumptions in computer vision algorithms. Additionally, assumptions are made as to the form of R. For example, Lambertian reflectance is often assumed, in which R is independent of V and depends only on the sum of the dot products of $\hat{n}$ with the illumination vectors $\{\vec{s}\}$.

Another aspect of the image formation model is the embedding of information about the sensor noise and the uncertainty in the reflectance model. The existence of reflectance model uncertainty means that $p_{d_0 f}(\vec{d}_0|\vec{f})$ (the conditional probability that we get the noise free sensory data $\vec{d}_0$ given that the observable object's surface normal vectors are $\vec{f}$) will not be a delta function, but will have some distribution that is somewhat spread out in the space of possible $\vec{f}$'s. This conditional density will be in general be very difficult, if not impossible, to specify exactly, and the usual procedure is to assume it to be a delta function (i.e. to assume that we know the reflectance model precisely).

Finally, we must take into account the distortion or noise added by the sensing device to the noise free data, $\vec{d}_0$. This will give us another conditional probability, typically not a delta function, $p_{d d_0}(\vec{d}|\vec{d}_0)$. The sensor noise process is usually taken to be additive (although for certain sensors, such as radar, multiplicative noise processes are more likely), and the conditional density is commonly assumed to be Gaussian. That is:

$$p_{d d_0}(\vec{d}|\vec{d}_0) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|M|}} e^{-\frac{1}{2}(\vec{d}-\vec{d}_0)^T M^{-1}(\vec{d}-\vec{d}_0)}$$

with M indicating the covariance matrix of the noise added to the sensor values, and n being the dimensionality of the data vector. Usually a further assumption that the elements of the noise added to the noise free data vector are independent is made, so that a univariate distribution can be used (i.e. $n = 1$), for each data element. If the surface reflectance conditional distribution, $p_{d_0 f}(\vec{d_0}|\vec{f})$ is taken to be a delta function $\delta(\vec{d_0} - \vec{f_0})$ then the image formation model is:

$$p(\vec{d}|\vec{f}) = p_{dd_0}(\vec{d}|\vec{d_0}) = p_{dd_0}(\vec{d}|L_\Omega[R(\hat{n}(\vec{x}); \vec{s}, V(\vec{x}))])_{T_{x \to y}}$$

If we do not assume a delta function for the surface reflectance distribution, then the determination of the image formation model is more difficult. In general we will have that:

$$p(\vec{d}|\vec{f}) = \int p_{dd_0}(\vec{d}|\vec{d_0}) p_{d_0 f}(\vec{d_0}|\vec{f}) d\vec{d_0}$$

If $p_{dd_0}(\vec{d}|\vec{d_0}) = p(\vec{d} - \vec{d_0})$ and $p_{d_0 f}(\vec{d_0}|\vec{f}) = p(\vec{d_0} - D(\vec{f}))$, where $D(f)$ is some arbitrary mapping of $\vec{f}$ then $p(\vec{d}|\vec{f})$ can be seen to be a convolution of p_{dd_0} and $p_{d_0 f}$. Hence, if we model the uncertainties in both the reflectance model and in the sensor noise model with a Gaussian distribution (probably not a very good model, but it will be difficult to come up with anything else that is really useful) then the resulting image formation conditional distribution will be Gaussian as well.

To finish our example, let us assume a Lambertian reflectance law, a single light source at infinity of unit intensity with direction $\hat{s}$, point sensors (in space and time), and assume that the same coordinate system is used for both the object and the sensor. Thus $d_0(\vec{x}) = \hat{n}(\vec{x}) \cdot \hat{s}$. Let us further assume that our uncertainty in our reflectance law (perhaps due to a random variation in the brightness of the light source) is represented by a Gaussian distribution, with variance σ_r^2, and that the sensor noise is additive and Gaussian with variance σ_n^2. We then have that $p(\vec{d}|\vec{f})$ is given by (via Theorem 2 of [110]):

$$p(d(\vec{x})|\hat{n}(\vec{x})) = \frac{1}{\sqrt{2\pi(\sigma_r^2 + \sigma_n^2)}} e^{-(d - \hat{n} \cdot \hat{s})^2 / (\sigma_r^2 + \sigma_n^2)}$$

An image formation model similar to this will be used in chapter 7, where we examine data fusional approaches to solving the shape from shading problem.

It should be evident from the above discussion that, like all constraints, the image formation model will usually be only approximately true at best, and may be completely incorrect in some situations.

2.3 THE PRIORS

The *a priori* probability $P(\vec{f})$ measures how likely a given $\vec{f}$ is *before* the measurement $\vec{d}$ is made. As pointed out in Chapter 1, many sensory information processing problems are ill-posed and assumptions (such as the smoothness constraint used in some stereo vision algorithms) are needed to solve them. In the Bayesian formulation these assumptions are incorporated in the *a priori* distributions $P(\vec{f})$ (often referred to as the "priors"). If a flat (or uniform, where every function $\vec{f}$ is equally likely) prior probability distribution is specified for an ill-posed problem then there may (depending on the form of the image formation model) be many possible solutions $\vec{f}$ which maximize (2.1).

In practice, the *a priori* distributions for the desired world information $\vec{f}$ are typically chosen to embody geometric, domain independent, assumptions. For example, in determination of the shape of an object's surface one can select the *a priori* distributions to favor surfaces that appear to be composed of elastic membranes or thin plates, possibly with discontinuities. Observe, however, that the formalism is rich enough to include domain specific constraints. For example, if a separate recognition process has identified an object as a human face then *a priori* knowledge about faces could be used to constrain the depth. The role of the *a priori* constraints is to select a unique solution out of the possibly infinite interpretations of the measured data. We will have more to say about the form of the priors in later chapters of this text.

2.3.1 THE SYSTEM MODEL

Often the world parameters that we are trying to determine from the data change over time. For example, the shape of the silhouette of an object may change as the object rotates. If we have some knowledge about how the system (the world or environment) is changing with time, we can use this knowledge to aid in the determination of the desired world parameters. The system model can be expressed in terms of a conditional probability density, $p(\vec{f}_t|\vec{f}_{t-1})$, where t and $t-1$ indicate successive time samples, and where $\vec{f}_{t-1}$ is the solution determined at the previous time step. This generalizes the notion of the prior model; we can replace the prior model $p(\vec{f})$, in the Bayesian formulation, by $p(\vec{f}_t|\vec{f}_{t-1})$, where $p(\vec{f}_1|\vec{f}_0) = p(\vec{f}_1)$ is the true *a priori* probability.

For some situations, the specification of a system model is straightforward. For example, the incremental nature of the depth estimation procedure of Matthies *et al* [106, 105] allows them to use knowledge about the motion of a moving sensor to model the change in the expected value of depth over time. In general, however, one does not have sufficient control over the manner in which the world parameters change to allow an accurate system model to be constructed. Even if a suitable form for the system model can be found, one may have a number of unknown parameters in the model. Without knowing these parameters, one cannot specify the conditional probability $p(f_t|f_{t-1})$. In such cases one has three options, 1) forego the use of a system model and rely solely on an *a priori* probability, 2) stick to a system model of a given form and use some external module to estimate the unknown model parameters, or 3) use an external module to determine both the form of the model and the values of the model parameters. The last case is better thought of in terms of having an external module that computes new *a priori* constraints (and not be tied to any model), that is, computing the $p(f)$ that is given to the Bayesian estimation process.

The last two alternatives listed above correspond to a form of

strongly coupled data fusion, that of *prior constraint adaption*, which will be discussed in detail in chapter 4. These techniques are distinguished by the alteration of the prior model (or of the system model) by an external process. For example, a motion analysis module could be used to track the motion of an object for use in determining the system model for an active (moving camera) depth estimation process, such as the one of Matthies *et al* [105].

2.4 BAYESIAN ESTIMATORS FOR $\vec{f}$

In order to determine the solution $\vec{f}$ using the Bayesian formulation, we must compute a suitable statistic of the *a posteriori* distribution of $\vec{f}$. The most obvious approach is to associate the "correct" solution with the $\vec{f}$ that maximizes the *a posteriori* distribution. This gives what is generally referred to as the *Maximum a Posteriori* (MAP) estimate of $\vec{f}$. Note that, in determining the MAP estimate, the *a priori* probability of the data, $p(\vec{d})$, since it does not depend on $\vec{f}$, acts merely as a normalization factor and does not affect the value of the estimate of $\vec{f}$.

If the prior model of $\vec{f}$, $p(\vec{f})$, is flat, or uniform, so that all solutions are equally likely *a priori* , the MAP estimate given above reduces to what is commonly known as the *Maximum Likelihood Estimate* (MLE) wherein one finds $\vec{f}$ that maximizes the likelihood $p(\vec{d}|\vec{f})$ of the data $\vec{d}$ given the candidate solution $\vec{f}$ [50].

Alternatives to the MAP or MLE estimates for $\vec{f}$ are often preferable. One possibility [50] is to find the minimum variance estimate (MVE), sometimes known as the minimum mean square error estimator (MMSE). This entails minimizing a quadratic functional:

$$J[\vec{f}] = \int_{\Omega} (\vec{f} - \vec{r})^T C(\vec{f} - \vec{r}) p(\vec{r}|\vec{d}) d\vec{r} \qquad (2.2)$$

where C is an arbitrary positive semidefinite weighting matrix, or *cost* matrix, and Ω is the space of possible solutions $\vec{f}$. The solution

to this minimization is, independent of $\mathcal{C}$:

$$\vec{f}^* = E[\vec{f}|\vec{d}] = \sum_{\vec{f}} P(\vec{f}|\vec{d})\vec{f}$$

That is, the solution is equal to the expected value (or mean) of $\vec{f}$ given the data $\vec{d}$, and is so often referred to as the *conditional mean* estimate. Other forms of loss functions besides the quadratic $(\vec{f}-\vec{r})^T C(\vec{f}-\vec{r})$ term in equation (2.2) result in the same estimate (subject to some weak restrictions on the form of $p(\vec{f}|\vec{d})$).

For linear image formation processes (i.e. when the transformation from $\vec{f}$ to $\vec{d_0}$ is linear, given by $\vec{d} = H\vec{f} + \vec{v}$, with H being a constant matrix and $\vec{v}$ a Gaussian white noise process with zero mean and covariance matrix R) and Gaussian distributions for $\vec{f}$ and $\vec{d}$, the minimum variance estimator generalizes many commonly used estimators. For example, in this case the MAP estimate is just the minimum variance estimate, and we get the MLE estimate if we assume that the distribution of $\vec{f}$ has infinite variance. We get the MMSE estimate if the distribution of $\vec{f}$ has zero variance. Furthermore, one can in the general case compute the conditional mean by applying a linear operator to the data. It can be shown [50] that the conditional mean is given by

$$\hat{f} = E[\vec{f}|\vec{d}] = (P_0^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \vec{d}$$

where P_0 is the *a priori* covariance matrix of $\vec{f}$. Note that the estimate of $\vec{f}$ is obtained by a linear transformation on the data $\vec{d}$.

As discussed in the earlier section on system models, the variable that we are trying to estimate, or its covariance, may change over time. If we have some knowledge of how the state variables (the world information that we are trying to estimate changes over time (e.g. from knowledge of sensor motion, or external information as to the motion of objects in the scene being sensed), we can come up with a system model that can be used in a *recursive filter* to provide better estimates. The recursive filter essentially derives a new "*a priori*" probability at each time step based on the previous estimates of the

state variable. A commonly used implementation of such a recursive filter is the Kalman filter [50]. The idea behind the Kalman filter[2] is as follows. Suppose that we have the following linear system and image formation models:

$$f_k = \Phi_{k-1} f_{k-1} + w_{k-1}$$

$$d_k = H_k f_k + v_k$$

where the indices k indicate the time (assumed to be discrete), Φ is the *state transition matrix* (i.e. this is our system model), H is the measurement transformation matrix, and where w and v are Gaussian white noise processes with covariance matrices Q and R respectively. The *a priori* information is embedded by setting $E[f_0] = \hat{f}_0$ and $E[(f_0-\hat{f}_0)(f_0-\hat{f}_0)^T] = P_0$ (i.e. the prior model is Gaussian with mean equal to the initial estimate, and covariance equal to the initial estimate of the covariance). The Kalman filter involves updating the covariance matrix [3] P of the variable to be estimated based on the image formation model, and updating the estimate, $\hat{f}$, of $\vec{f}$ based on the current estimate of the covariance P_k, the image formation model, and the system model. The update equations are given below (the derivation of these can be found in [50])

$$\hat{f}_k = \Phi_{k-1}\hat{f}_{k-1} + K_k(\vec{d_k} - H_k\Phi_{k-1}\hat{f}_{k-1})$$

$$P_k = (I - K_k H_k)(\Phi_{k-1} P_{k-1} \Phi_{k-1}^T + Q_{k-1})$$

The matrix K_k is called the Kalman filter gain and is given by

$$K_k = P_k H_k^T R_K^{-1}$$

It is important to realize that the Kalman filter and other linear estimators rely on the assumption that the image formation model is linear and the distribution of the sensor noise and the *a priori*

[2] The *discrete* Kalman filter is described here. The extension to the continuous version is detailed in [50].

[3] P is sometimes referred to as the error covariance, since the estimation error at time k, $(\hat{f}_k - \vec{f})$ has a Gaussian distribution with mean zero and covariance P_k.

estimate is Gaussian. It is more usually the case, however, that either the image formation process is non-linear or the distribution of $\vec{f}$ is non-Gaussian, or both. In these situations the above statements are not true and there are many possible estimates of $\vec{f}$ that are distinct from the conditional mean. In addition, the computation of the conditional mean requires some form of nonlinear operation, such as the Extended Kalman Filter [50]. In the next chapter we will investigate energy function approaches derived from mean field theory considerations for estimating the conditional mean.

One must be careful, however, in coming up with a suitable estimator for $\vec{f}$ when one has non-Gaussian distributions. It will often be the case that the variance of such distributions do not exist, or cannot be defined. In such situations one can not construct a meaningful minimum variance estimate. For example, some distributions, such as the Cauchy distribution, have an infinite variance. A more serious problem can be seen when one looks at spaces F of possible solutions $\vec{f}$ for which there is no suitable underlying metric to use in computing the mean and variance. For example, consider the space F to be composed of apples, oranges and bananas. We can construct a conditional density which expresses the probability that we are sensing the various elements of F given our sensory data, but it makes no sense to talk about the mean or variance of this conditional density. In the situations where the variance of the *a priori* distribution does not exist, we must use estimators other than the conditional mean or minimum variance estimator. The MAP estimate will always be defined, but other estimators based on "pseudo-variance" measures may be used.

2.5 BAYESIAN DETECTION AND EXTRACTION SYSTEMS

The methods described in the previous section are all concerned with the estimation of some continuous (usually) valued parameter

based on the data and *a priori* constraints. In the language of statistical decision theory [109, 154] this is called *parameter estimation or extraction,* and the Bayesian estimators are those which minimize the *average risk.* Here, *risk* refers to a simple form of a *cost,* or *loss* function, F= $C(S,\gamma)$, where F is the cost of making the decision γ, independent of the decision rule or estimation process, when the actual parameter or signal is S. For a particular decision rule $\delta(\gamma|\vec{d})$ the average risk is given by

$$R(p(S),\delta) = \int_\Omega p(S)dS \int_\Gamma p(\vec{d}|S)d\vec{d} \int_\Delta C(S,\gamma)\delta(\gamma|\vec{d})d\gamma$$

In the above, $p(S)$ is the *a priori* probability of the signal S, Ω is the space of all possible signals, Γ is the space of all possible data values, Δ is the space of possible decisions, and $\delta(\gamma|\vec{d})$ is the decision rule to be used (i.e. given $\vec{d}$, what decision γ do we make?).

The optimal Bayes estimator, or decision rule, can be defined as the one which minimizes the average risk $R(p(S),\delta)$ for the given $p(S)$. It is evident that the decision rule that we obtain will depend on the form of the cost function that we choose. It can be shown (e.g. Middleton [109], chapter 21.2-2) that the optimal Bayes estimator for a quadratic cost function and Gaussian image formation models is the conditional mean (as was stated in the previous section). In the same reference it is shown that appropriately defined "constant" cost functions result in optimal Bayes estimators that are Maximum Likelihood Estimators.

Often one has a problem wherein the space of possible solutions contains a finite number of discrete elements. In these types of problems one is concerned with "deciding" which of the finite number of possibilities is the correct one, rather than estimating the value of some continuous parameter. This is often referred to as a *detection* process, wherein the presence or absence of a particular signal is determined (and which can be generalized to multiple types of signals). The detection process should be contrasted with the extraction processes described above, which are concerned with estimating some property of the detected signal. The idea behind the process is the

same, however, and the optimal Bayes decision rule is that which minimizes the average risk. The case of the detection process is simpler than the extraction process as the signal and decision spaces now contain only a finite number of discrete elements so that the integrals in the average risk turn into summations. In addition the cost functions are usually much simpler. For example, binary detection problems with constant, combinatorial, cost functions, the decision process reduces to testing whether a particularly defined *likelihood ratio* exceeds a threshold or not. [109]. An example of such a binary detection process is given in chapter 8.

The techniques for solving detection problems have been primarily used for communication systems design [109, 154], but examples of similar problems abound in sensory systems, especially at the higher cognitive levels. These problems are typically either segmentation problems, wherein an image is partitioned into a number of distinct regions, and classification problems, wherein areas of the image are classified as belonging to one of a set of different classes. A representative example is illustrated in chapter 8 where we consider the problem of segmenting an image into regions of specular and Lambertian reflection.

2.6 THE BAYESIAN CONTROVERSY

The innocuous looking formula (2.1), in particular the *a priori* term $P(f)$, conceals a number of fundamental philosophical and scientific issues related to the origins of knowledge, the foundations of statistics and theories of perception.

The Bayesian approach is an interesting one as it has had a long and stormy history with regard to the science of the acquisition of information in general (for a short history and discussion of the Bayesian approach see the monograph by Weber [158]). Early in the growth of statistical science the commonly accepted view of inference (or of information acquisition) was that it must be based

only on observations or measurements and not on prior information or beliefs. This objective viewpoint associates the probability of the occurrence of an event with the limit of the relative frequency of its occurrence in some specified class of events as the number of trials goes to infinity. The objective approach to probability has been championed by Von Mises [152], Reichenbach [130], and Carnap [29] among others. As pointed out by Weber [158] the objective view suffers from

> "...the impossibility of determining with certainty the limit of a relative frequency; thus the estimate of a probability may require repeated revision."

Many statisticians now acknowledge that our interpretations of statistical experiments should depend, to some extent, on our *subjective* or *personal* beliefs about the hypothesis being tested. For example (taken from Savage [137]), suppose a drunken friend bets that he can predict the toss of a coin ten times in succession. The chance of him doing this correctly by random choice is 2^{-10}, yet we are likely to attribute any success he may have to luck rather to skill. On the other hand if a musician claims to be able to distinguish between a page of Haydn's score from a page of Mozart's score and is successful in ten trials we are likely to believe him. The difference lies in our *a priori* expectations of how likely the claims are. The basic idea behind the subjective view of probability is that *the data should be used only to tell us how to modify our opinions (or subjective probabilities), and not to decide which opinions are justifiable* [138, 158]. The subjective approach is usually criticized on the grounds that it cannot justify the *a priori* opinion that it requires, and that it is often unclear how to determine the prior probability distributions. Some statisticians have therefore argued against using priors, unless there is a rigorous or objective way to select them. This has lead to divisions in the statistical community between Bayesians and Non-Bayesians, Subjectivists and Frequentists. Berger [11] and Weber [158] examine these issues in more detail.

Bayes theorem is most useful in the framework of subjective probabilities, and it is for this reason that many statisticians object to the use of Bayesian methods. It is because of the difficulty in specifying the *a priori* constraints and even the image formation model that some statisticians consider Bayesian methods to be useless. What we have been calling the image formation model is generally assumed to be objectively determined, but in practice is usually based on subjective opinion, although these opinions may be generally regarded as valid. The *a priori* distributions, on the other hand, are usually exceedingly subjective, and differ widely between individuals. Referring to the taxonomy of constraints talked about in Chapter 1, we could say that physical constraints tend towards objectivity (and hence the image formation models, which are usually based on physical constraints, are thought of as rather objective), while natural constraints are more subjective. Artificial constraints are purely subjective, or personal, as different observers will have different high level expectations of what they are looking at, depending on their current state and previous experience.

Are the criticisms of subjective methods, and hence of the Bayesian approach, really valid? This is still open to debate, but strong arguments can be made in favor of the subjective approach. One of the criticisms of the subjective view is the lack of justification for the priors. However, as pointed out by Weber [158], the *a priori* opinions or subjective probabilities need not be justified, since

> "...in fact subjective theory does not contend that opinions about probability are uniquely determined and justifiable. That is, subjective probability does not correspond to a "rational" belief, but rather to an effective personal belief."

This says, essentially, that the result of an experiment from the subjective probability point of view is to alter whatever prior opinions the experimenter may have held, and that this result is therefore just as subjective as the prior belief. The experimenter may have a

personal belief that his prior opinions are completely objective (i.e. taken from some venerable, but up-to-date, textbook on science), but these opinions are nevertheless subjective, as other experimenters may have different opinions, and, as was mentioned earlier, there is no way that truly objective probabilities can be obtained in finite time. Thus, when one understands that no piece of information can ever be objectively determined, and that all information is thereby subjective, the arguments against subjective methods vanish. The plight of the objectivist is summarized in the following quote by I.J. Good [56]

> "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

How does the Bayesian controversy manifest itself in sensory information processing systems? We saw in Chapter 1 that, because of the ill-posed nature of the sensing process, sense data alone is insufficient to solve for world information. We need extra information, in the form of *a priori* constraints, to be able to uniquely invert the world-image mapping. One could argue, as Gibson does [54], that this statement is erroneous and that the sense data alone does contain enough information to obtain a unique inversion (e.g. by using the "sensory flow"). However, such arguments are equivalent to the arguments in support of objective probabilities, and fail for the same reasons. For the application of sensory information processing (in biological or artificial active systems) the subjective approach is most natural. Gibson's own theory of affordances [54], which formed part of his ecological approach to sensing, is perhaps one of the most compelling advocates of the use of the subjective approach. The affordance based viewpoint says that organisms perceive what is useful to the activities that they wish to undertake. This implies that the perceptual process is a subjective one in that different types of organisms will interpret the same sensory data in different ways, according to their needs.

The *a priori* constraints that an organism uses in performing sensory tasks, if we follow the affordance paradigm, are defined by the aspects of the environment that are of importance to the organism. How are these priors determined? Clearly, subjective knowledge of both the environment and the activities of the organism are required to specify the domain for which the organism's sensory activity is valid. The knowledge of the activity is used to determine which aspects of the environment are of importance and which can be ignored, or safely misinterpreted. For example, for an assembly robot looking for a screwdriver in a field of flowers, it is safe to misinterpret the flowers as engine gaskets, as neither interpretation is of importance to the robot's activity. This can lead to trouble, of course, if the the robot misinterprets irrelevant objects as objects of importance to its activities. For example, some birds have been known to treat strange objects, such as bottles and rocks, that have been put into their nests by mischievous scientist, as their own eggs. This is due to misinterpretations based on invalid affordance motivated assumptions. These invalid assumptions invariably lead to inappropriate activities. Obviously, these assumptions are not always invalid, and in fact, for most of the history of birds, they have not had to contend with inquisitive scientists in their domain. Their sensory systems have been designed under the assumption that the sorts of tricks that avian psychologists play will never be encountered, and these systems generally work well in most situations that a bird is likely to find itself in.

The assumptions or *a priori* constraints dealing with the activities that an organism undertakes are most often innate or hardwired. The priors dealing with the structure of the external world or environment are obtained, in part, as the product of scientific endeavor (i.e. more or less objectively, or what is usually construed as objectively). These are the physical constraints of chapter 1. These constraints are usually so tightly coupled to the very structure of the algorithm, however, that they usually don't appear explicitly as *a priori* constraints, but rather show up in the image formation model. Other constraints are not so tightly tied to physical processes but are more subjective in nature. They still somewhat reflect

the underlying physical (objective) order, however. In the context of vision modules Marr [97] refers to these types of constraint as *natural constraints* and suggests they can be determined by psychophysical experiments and detailed analyses of the module. He felt that the process of determining the form of these constraints is what made Computational Vision a science. In general, the natural constraints form the bulk of the *a priori* assumptions used in sensory information processing algorithms, and it is on the examination of these constraints that any answer to the philosophical question of the propriety of the Bayesian formulation must rest. Acceptance of the subjective approach to information processing somewhat simplifies the problem of determining suitable *a priori* constraints for a given information process task. We choose these constraints based on our subjective opinions, colored by the activities that we intend to engage in, of the likelihood of possible solutions. This is an improvement (in the sense of requiring less computation) over the objective approach since few experiments need to be performed to arrive at our opinions (this number can be quite small and the Bayesian approach will still happily work away), and the priors can be chosen to reflect the purpose of our sensing activity (as in the affordance paradigm).

One might ask what would be the result of a subjective analysis if one observer made many experiments, or what would the result be if many different observers observed the same experiment? These two questions are key to the major theme of this book, *data fusion*. With regard to the first question it is clear that as the number of independent measurements increase, the conclusions become independent of the prior expectations. In this case the subjective and objective approaches lead to similar results. Hence, if we can *fuse* the results of many different experiments, e.g. many sets of sensory data, we can avoid many of the drawbacks of the subjective approach. This, of course, assumes that the initial priors are somewhat consistent with the measurements (else the measurements will be ignored as being faulty), and that the environment is not changing (so that the measurements refer to the same process, and can hence be meaningfully fused).

The second question relates to the use of what is known as *multisubjective priors* [36]. That is, we have many observers of an experiment, each with their own subjective priors or personal opinions regarding the expected outcome of the experiment. Definetti [36] suggests that one can create decision rules that are acceptable to many observers, even in light of their differing *a priori* beliefs. Such a decision process would imply a fusion operation, wherein what is fused are not multiple sets of data, but rather multiple sets of *a priori* constraints. The result would still be subjective, as it was based on subjective priors, but may be considered more objective due to the consensus of the multiple observers. Of course this depends on the veracity of the observers and the objective validity of their opinions. In this text we apply this concept of "constraint fusion" to sensory information processing problems and come up with new methods for fusing or integrating multiple sensory modules. In contrast to the usual approach to integration of modules, we look at integrating the *a priori* constraints of the modules as well as the data that these modules produce. In other words we look at the sensor integration process as one of "one experiment, multiple observers", and derive techniques for integrating the observation processes (and not just the observations). Our methods will include a new class of fusion algorithms, which we call strongly coupled algorithms, which involve different sensory modules affecting the operation of other sensory modules. In the multisubjective view this corresponds to one observer altering the opinions of another, and thereby reaching consensus. In chapter 8 we describe an example of this concept wherein we have two observers, each with a different image formation model, that view the same data. The fusion process involves deciding which image formation model to "believe" in extracting the desired world information (in this case the shape of an object) from the data.

In the next chapter we will look at a class of information processing techniques that are related to the Bayesian approach. These methods are based on the formulation of information processing tasks as the minimization of energy functionals. We provide a number of different approaches that are suitable for sensory information processing problems. We will also investigate a number of ways in which

the *a priori* constraints used in the Bayesian or energy function approaches can be formulated, including parametric constraints and minimum length codes.

2.7 CHAPTER SUMMARY

- Bayes rule provides an intuitively satisfying means of embedding constraints into an information processing task. The principle aspect of this approach is the probabilistic representation of constraints. Solutions to the world-image mapping are assigned probabilities corresponding to their likelihood with respect to the constraints.

- The Bayesian formalism has two primary components, the image formation model and the prior model. The image formation model represents the conditional probability of a given data value arising from a particular world configuration. It captures the process by which data is generated from different world configurations. The prior model represents the *a priori* probability of the particular world configuration, and corresponds to the expectations of what configurations structures in the world are likely to occur.

- There are a number of different statistics that can be used to determine a solution (world configurations) from the *a posteriori* probability density. Some of the most useful are the Maximum A Posteriori (MAP) estimator, which corresponds to the mode of the *a posteriori* density, the Maximum Likelihood estimator (MLE) which is obtained from the MAP estimate when the *a priori* density is flat, and the Minimum Variance estimator (MVE), which corresponds to the mean of the *a posteriori* density.

- Statistical Decision theory can be used to derive alternative estimators. These are based on minimizing *average risk*, where

the risk of an estimate or solution is based on some cost defined using prior knowledge.

- We treat all probabilities used in our approach as subjective. They are dependent on the activities that the organism that uses them is involved in. Objective probabilities are equivalent to universal constraints, which can never be specified exactly.
- The concept of data fusion, where we use information from multiple sources to reduce dependence on *a priori* constraints, can be thought of as specifying multisubjective probabilities, where each module has its own set of prior expectations, and these are combined in some fashion to obtain a consensus subjective expectation during the fusional process.

Chapter 3

Information Processing Using Energy Function Minimization

A good deal of recent research into sensory information processing algorithms has focussed on the so-called energy function minimization, or regularization, approach to inverting the world-image mapping. In this chapter we review the application of this paradigm to early vision and its incorporation within the Bayesian framework. We will concentrate in this section on using this formulation to impose smoothness constraints on the solutions; other constraints can be similarly imposed, as will be described in later sections. A large part of this chapter will be concerned with new approaches to formulating energy function based algorithms using techniques borrowed from statistical mechanics.

The idea of using energy functions to impose smoothness constraints in early vision has been very influential (similar applications to higher level vision by Rosenfeld, Hummel, and Zucker [135] and to motion correspondence by Ullman [150] are of note as well) and we

mention here a few early examples. Ikeuchi and Horn [72] describe a theory of shape from shading using a variational principle and Ikeuchi [73] uses a similar technique for shape from texture. In this work Ikeuchi describes how his technique is able to impose smoothness constraints on the object and draws the analogy with imposing constraints in Artificial Intelligence. Horn and Schunk [65] applied these techniques to optical flow. Hildreth [60] used a similar method to solve the aperture problem for motion based on zero crossing contours. Grimson [57] uses a related approach to interpolate a surface through sparse stereo data. Terzopoulos [148] extended Grimson's work using more sophisticated techniques.

Poggio and Torre [125] noted the similarity of these methods to a branch of mathematics called regularization theory [149] and proposed a unified framework for vision algorithms. The switch from energy functions to regularization required a few changes in terminology (for example "smoothness term" becomes "regularizer") but no fundamental alterations to the theory were necessary. Both of these methods had an important property that was both a weakness and a strength; they imposed continuous solutions and smoothed over discontinuities. Regularization theory [149] indeed required that the solution to a problem depended smoothly on the data. Inserting a discontinuity in the solution would require a yes/no decision, and hence could not depend continuously on the data. Such decisions would have to be made by a separate process; we discuss one such process in the next section.

The inability to deal with discontinuities however had important practical advantages, as the energy functions tended to be convex and not have local minima. Thus they could be minimized by simple methods such as gradient descent. More sophisticated techniques could be used to speed up the convergence. For example, Terzopoulos [148] adapted a multi-grid algorithm due to Brandt [21] to the task of minimizing energy functionals.

As a simple example of the regularization technique consider the problem of smoothing a one-dimensional image $d(x)$ for the purposes

of taking derivatives [126]. Regularization theory involves minimizing a functional

$$E[f(x)] = \int \{f(x) - d(x)\}^2 dx + \lambda \int \{Lf(x)\}^2 dx$$

with respect to the function $f(x)$. The first term is the *data term* (sometimes referred to as the consistency term or as the fidelity term) which requires that the smoothed solution should be close to the data. The second term, the *regularizer*, imposes smoothness on the solution by a suitable differential operator L. The constant λ determines the relative importance of the two terms. In some situations an optimal value of λ can be estimated [47] using cross-validation techniques [153]. The functional is positive definite and quadratic in $f(x)$. Therefore it is convex and hence has a unique minimum [35].

In most implementations the data is specified on a regular lattice and the energy function is discretized. If we choose the smoothness operator to be $L = d/dx$, corresponding to fitting the data with an elastic membrane [35], the energy functional becomes

$$E(f_i) = \sum_i \{f_i - d_i\}^2 + \lambda \sum_i \{f_{i+1} - f_i\}^2 \qquad (3.3)$$

The energy function can be minimized by directly solving the Euler-Lagrange equations which, in the discrete case, are equivalent to solving the following system of linear equations

$$\frac{1}{2}\frac{\partial E}{\partial f_i} = f_i - d_i + \lambda\{2f_i - f_{i+1} - f_{i-1}\} = 0, \quad for \ \ all \ \ i.$$

Alternatively the minimum can be found by steepest descent techniques [92]. The simplest method corresponds to iterating

$$f_i(t + \delta t) = f_i(t) - K\frac{\partial E}{\partial f_i} \qquad (3.4)$$

where K is a constant. It is straightforward to check that the energy decreases monotonically as we iterate (3.4), thus the system converges to the minimum of $E(f_i)$.

An alternative approach to minimizing the energy functions was introduced into vision by Yuille and Grzywacz [165][1] based on work described in [40]. Suppose we have a sparse set of data points $d(x_i)$ through which we want to interpolate. We can choose the smoothness constraint (regularizer) to be

$$\sum_{n=0}^{\infty} a_n \int \left(\frac{d^n f(x)}{dx^n}\right)^2 dx$$

where the a_n are constants.

Then the energy function becomes

$$E[f(x)] = \sum_i \{f_i - d_i\}^2 + \lambda \sum_{n=0}^{\infty} a_n \int \left(\frac{d^n f(x)}{dx^n}\right)^2 dx$$

This can be written as

$$E[f(x)] = \sum_i \int \delta(x - x_i)\{f(x) - d_(x)\}^2 dx + \lambda \sum_{n=0}^{\infty} a_n \int \left(\frac{d^n f(x)}{dx^n}\right)^2 dx$$

where $\delta(x)$ is the Dirac delta function.

The Euler-Lagrange equations [35] are

$$\sum_i \{f_i - d_i\}\delta(x - x_i) = \lambda \sum_{n=0}^{\infty} a_n (-1)^{n+1} \frac{d^{2n}}{dx^{2n}} f(x) \qquad (3.5)$$

It then follows [165] that the solutions are of the form

$$f(x) = \sum_i \alpha_i G(x - x_i) \qquad (3.6)$$

where $G(x)$ is the Green's function of the differential operator $\sum_{n=0}^{\infty} a_n (-1)^{n+1} \frac{d^{2n}}{dx^{2n}}$, i.e.

$$\sum_{n=0}^{\infty} a_n (-1)^{n+1} \frac{d^{2n}}{dx^{2n}} G(x) = \delta(x)$$

[1] Aloimonos and Schulman have implemented a version of this theory.

and α_i satisfy (after substituting (3.6) into (3.5))

$$\sum_j \{\delta_{ij} + \lambda G(x_i - x_j)\}\alpha_j = d_i$$

Thus minimizing the energy function is reduced to inverting a matrix equation for the α_i.

This approach highlights an important point on the choice of priors emphasized in [165] for motion perception. When doing interpolation, or estimating optical flow, it is often desirable that the influence of nearby data points is more important than that of far away points. Mathematically this is equivalent to requiring that the Green's function $G(x)$ of the smoothness operator decreases with x. For many common choices of regularizer, for example the choice $L = d/dx$, the Green's function increases rather than decreases with x. In [165] it is shown that by suitable choice of the coefficients a_n the Green's function can be chosen to be a Gaussian, hence having the correct falloff properties. The details of this are given later in section 3.4.

The energy minimization/regularization approach can be incorporated directly into the Bayesian framework using the Gibbs distribution [119]. To each possible configuration of the f_i we can define a probability

$$P(f_i) = \frac{e^{-\beta E(f_i)}}{Z} \qquad (3.7)$$

where β is a constant (whose interpretation we will discuss in section 3.3) and Z is a normalization constant.

There are other ways of associating probability distributions to energy functions. The Gibb's distribution is usually preferred since, as well as its ubiquitous use in physics, it turns quadratic energy functions into Gaussian probability distributions, and turns sums of energy terms into products of probability distributions. To see this consider the energy function defined in (3.3). This corresponds to a

probability distribution consisting of products of Gaussian functions

$$P(\vec{f}|\vec{d}) = \frac{1}{Z}\prod_i e^{-\beta(f_i-d_i)^2}\prod_i e^{-\beta\lambda(f_{i+1}-f_i)^2} \quad (3.8)$$

The first term $\prod_i e^{-\beta(f_i-d_i)^2}$ corresponds, when suitably normalized, to the $P(\vec{d}|\vec{f})$ term in (2.1). The second term $\prod_i e^{-\beta\lambda(f_{i+1}-f_i)^2}$ is the prior distribution of the f_i. Note that the prior distribution is, strictly speaking, not well defined since it only specifies the relative values of the f_i and not their absolute values. That is, $p(\vec{f}) = p(\vec{f}+\alpha)$ for any constant α. There is also a problem due to the fact that $p(\vec{f})$ as we have defined it is not normalizable, and is therefore strictly not a density. The maximization of the conditional density, however, does not depend on the normalizability of $p(\vec{f})$ so we will retain our formulation of $p(\vec{f})$.

It follows directly from (3.7) that minimizing an energy function $E_{\vec{f},\vec{d}}$,

$$\vec{f}_{opt} \leftarrow min_{\vec{f}}\left[E_{\vec{f},\vec{d}}\right]$$

is equivalent to maximizing the corresponding probability distribution

$$\vec{f}_{opt} \leftarrow max_{\vec{f}}\left[e^{-\beta E(\vec{f},\vec{d})}\right]$$

Thus all problems whose solution can be obtained by minimizing an energy function can be expressed in the Bayesian framework.

3.1 MARKOV RANDOM FIELDS

If we use the Gibb's distribution on discretized energy functions, such as are obtained from discrete sensing elements, we obtain a special mathematical structure known as a Markov Random Field (or MRF).

A MRF [161] consists of a probability distribution over a set of variables $\{f_i\}$ such that the probability of a specific variable f_i depends only on the state of its neighbors. More precisely, we can define a neighborhood N_i such $P(f_i|f_j, j\epsilon N_i) = P(f_i|f_j, \ for\ all\ j)$. For the energy function (3.3) we see, from (3.8), that the Gibb's distribution satisfies this condition with a neighborhood structure $N_i = \{i-1, i+1\}$. The MRF structure therefore incorporates all the energy functions we have discussed so far.

One of the most influential formulations of vision problems in terms of MRFs has been the work of Geman and Geman [51] on image segmentation (see also the closely related work of Blake [14] discussed later this section). The aim of their work was to smooth images except at places where the image values were rapidly changing. In the one dimensional case this can be done by defining an energy function

$$E(f_i, l_l) = \sum_i\{f_i - d_i\}^2 + \lambda\sum_i\{f_{i+1} - f_i\}^2(1-l_i) + \mu\sum_i l_i \quad (3.9)$$

where $\{l_i\}$ is a binary line process field (i.e. a field of elements that can be in one of two possible states). When switched on, e.g. $l_i = 1$, a line process element breaks the smoothness constraint on $\vec{f}$ between i and $i+1$. For each line process element that is turned on we add a cost μ to our energy.

Using the Gibb's distribution on $E(f_i, l_i)$ we obtain two coupled Markov Random Fields $\{f_i\}$ and $\{l_i\}$. The goal is find the most probable configuration of $\{f_i\}$ and $\{l_i\}$ given this distribution or, alternatively, to minimize $E(f_i, l_i)$ with respect to $\{f_i\}$ and $\{l_i\}$ simultaneously. Note that one could conceive of a system that would compute an $\{f\}$ and an $\{l\}$ independent of each other, that is, have separate image smoothing and image edge detection modules. The coupled MRF process is different in that the computation of the smoothing and edge detection modules are not independent of each other but are *strongly coupled.* This form of cooperative computation is a key aspect of the approach to data fusion expounded in this book. We will have more to say about strong coupling of sensory mo-

dules in the next chapter, and will give examples of its application in later chapters.

The minimization process is not straightforward as $E(f_i, l_i)$ will usually have many local minima. Geman and Geman [51] attempted to find the global minimum using simulated annealing [79]. This technique involves using a Monte Carlo algorithm [107] to bring the system to thermal equilibrium at temperature $1/\beta$. The temperature is gradually reduced to 0, following an annealing schedule, and the system can be shown to converge to the lowest energy state. Unfortunately, lower bounds on the rate of change of the temperature from the annealing schedule are very low and cause the algorithm to often be very slow. In practice the algorithm often gives good results if the annealing schedule is not followed, but the guarantee of correct convergence is lost.

For image segmentation the coupled Markov Random Field approach is a significant improvement on the regularization method described in the previous section. The line process fields prevent smoothing across discontinuities in the image. They can also be used as part of a framework for data fusion [123], as is described in the next chapter.

3.2 ENERGY FUNCTIONS WITH MATCHING ELEMENTS

This section describes the minimization of energy functions which include terms associated with the matching of image elements. A number of early vision modules, such as stereo or long-range-motion, involve matching features. Ullman [151] proposed a theory of long-range-motion correspondence which assumes that the process can be modeled as a matching problem between consecutive time frames. More precisely, suppose there are N features at positions $\vec{x}_i$ in the t'th time frame and M features at position $\vec{y}_a$ in the $t+1$'th frame.

We can define binary matching elements V_{ia} such that $V_{ia} = 1$ if the feature at $\vec{x}_i$ matches the feature at $\vec{y}_a$ and $V_{ia} = 0$ otherwise. Ullman's minimal mapping theory [150] proposes choosing the V_{ia} so as to minimize an energy function

$$E(V_{ia}) = \sum_{ia} V_{ia} \, |\vec{x}_i - \vec{y}_a|^2$$

where the minimization is taken over the set of V_{ia} such that each point has at least one match but the total number of matches is the least possible (i.e. $max(N, M)$); this is called the *cover principle* [150].

At first sight this energy function seems rather different from the ones we discussed in previous sections since it does not involve a smoothness constraint. However, as we now show (following [165]), it can be related to such theories.

In [165] it proposed that the problem be treated in terms of finding a smooth velocity field $\vec{v}(\vec{x})$. This involves defining a matching term and a smoothness term. The matching term is

$$E_{match} = \sum_{i,a} V_{ia}((\vec{y}_a - \vec{x}_i) - \vec{v}(\vec{x}_i))^2 \qquad (3.10)$$

where we have normalized the time difference between frames to be unity. If $V_{ia} = 1$, then the point at $\vec{x}_i$ in the first frame travels to point $\vec{y}_a$ in the second, and therefore it corresponds to a velocity of $\vec{y}_a - \vec{x}_i$ (remembering the time normalization). The velocity field at $\vec{x}_i$ is therefore constrained by this value. If $V_{ia} = 0$ then the points are unmatched and no contribution is made to the energy. To obtain the full energy function we add a smoothing term for the velocity field as before to obtain

$$E(V_{ia}, \vec{v}(\vec{x})) = \sum_{i,a} V_{ia}((\vec{y}_a - \vec{x}_i) - \vec{v}(\vec{x}_i))^2 + \lambda \int \sum_{m=0}^{\infty} c_m \, (D^m \vec{v})^2 \quad (3.11)$$

We minimize this function over V_{ia} and $\vec{v}(\vec{x})$ simultaneously. Thus, the smoothness requirement directly affects the matching in this

case. This is another example of the strong coupling of information processing modules (in this case feature matching and velocity field smoothing). The V_{ia} must be constrained so that points typically make exactly one match. There are two possible ways to do this; (i) to use the cover principle [150] and require that all points have at least one match, or (ii) to incorporate a cost in the energy function to bias against too many or too few matches, for example $E_c = (\sum_a(\sum_i V_{ia} - 1)^2 + \sum_i(\sum_a V_{ia} - 1)^2$.

We now show that the energy function can be transformed into a form which includes minimal mapping theory as a special case. The key observation is that $E(V_{ia}, \vec{v}(\vec{x}))$ is quadratic in $\vec{v}(\vec{x})$. The Euler-Lagrange equations for $\vec{v}(\vec{x})$ will, therefore, be linear and can be solved for as a function of V_{ia}. We now express $E_M(V_{ia})$ in terms of the V_{ia} only.

We first note that the horizontal and vertical components of the velocity field in (3.11) do not interact so that we can treat each component separately. Henceforth, we consider the horizontal component only. The Euler-Lagrange equations become

$$\lambda \sum_{m=0}^{\infty} c_m \nabla^{2m} v(\vec{x}) = -\sum_{i,a} V_{ia}(v(\vec{x}_i) - d_{ia})\delta(\vec{x} - \vec{x}_i) \qquad (3.12)$$

where $d_{ia} = y_a - x_i$ and δ is the Dirac delta function. The c_m are given by $c_m = \sigma^{2m}/(m!2^m)$. This choice ensures that the Green's function of the operator $\sum_{m=0}^{\infty} c_m \nabla^{2m}$ is the Gaussian $G(\vec{x}, \sigma)$ [165]. In other words

$$\sum_{m=0}^{\infty} \nabla^{2m}\sigma^{2m}/(m!2^m)G(\vec{x}, \sigma) = \delta(\vec{x}) \qquad (3.13)$$

Using (3.13) we see that the solution to (3.12) is of form

$$v(\vec{x}) = \sum_i \beta_i G(\vec{x} - \vec{x}_i). \qquad (3.14)$$

Substituting this back into (3.12) we get

$$\lambda \sum_i \beta_i \delta(\vec{x} - \vec{x}_i) = -\sum_{i,a} V_{ia}(v(\vec{x}_i) - d_{ia})\delta(\vec{x} - \vec{x}_i)$$

Equating coefficients of the delta functions gives us a set of equations

$$\lambda\beta_i = -\sum_a V_{ia} v(\vec{x}_i) + \sum_a V_{ia} d_{ia}$$

Now, we can use $\sum_a V_{ia} = 1$ and substitute for v from (3.14)

$$\lambda\beta_i = -\sum_j G_{ij}\beta_j + \sum_a V_{ia} d_{ia}$$

where $G_{ij} = G(\vec{x}_i - \vec{x}_j)$. Using the Einstein summation convention (i.e. summing is done over repeated indices), and the Kroenecker delta function we find

$$(\lambda\delta_{ij} + G_{ij})^{-1}\beta_j = \sum_a V_{ia} d_{ia} \qquad (3.15)$$

Now we substitute back for β_i into the energy function. This gives, using the summation convention,

$$E(V_{ia}) = \lambda^2\beta_i\beta_i + \lambda\beta_i G_{ij}\beta_j$$

Using (3.15) gives

$$E(V_{ia}) = \lambda(\sum_a V_{ia} d_{ia})(\lambda\delta_{ij} + G_{ij})^{-1}(\sum_b V_{bj} d_{bj})$$

In the limit as $\sigma \mapsto 0$ we get that $\lambda\delta_{ij} + G_{ij} \mapsto \delta_{ij}/(2\pi\sigma^2)$ and thus

$$E_{limit} = 2\lambda\pi\sigma^2(\sum_a V_{ia} d_{ia})^2$$

This has the same minimum as a function of V_{ia} as the minimal mapping theory, which has an associated energy function given by

$$E_{mm} = \sum_{ia} V_{ia} |d_{ia}| \qquad (3.16)$$

This limit is also found if we let $\lambda \mapsto \infty$.

By comparing (3.10) with (3.16) we see that minimal mapping can be thought of as a version of this theory as $\sigma \mapsto 0$. This

corresponds to not smoothing the velocity field. Hence the Yuille-Grzywacz theory imposes more smoothness on the velocity field than does minimal mapping and so is better suited for dealing with motion capture phenomena.

The energy functions we have used here can be naturally given a probabilistic interpretation by using the Gibb's distribution as described earlier.

3.3 STATISTICAL MECHANICS AND MEAN FIELD THEORY

In this section we introduce some techniques from statistical physics [119] principally mean field (MF) theory, which we can use for analyzing MRF models. These techniques have been applied to vision by Geiger and Girosi [46], Yuille [166], Marroquin [102], and by Geiger and Yuille [49].

As mentioned in section 3.4, the conditional mean or minimum variance estimate is often to be preferred to the maximum a posteriori estimate. In this section we describe (following [48]) how the conditional mean estimate can be calculated using techniques from statistical physics.

An important concept in statistical mechanics is the partition function Z, defined for a probability distribution $P(f,l) = e^{-\beta E(f,l)}$ as

$$Z = \sum_{f,l} e^{-\beta E(f,l)}$$

Many statistical properties of the fields $\{f\}$ and $\{l\}$ can be calculated from Z. The partition function can be thought of as analogous to the generating functions used in conventional probability theory.

MF theory gives methods for obtaining fast deterministic algorithms, that make use of the richness of the statistical formulation, to find the MF solution. This solution corresponds to the mean of the probability distribution and will not necessarily correspond to the solution that minimizes the energy, especially if that state is very isolated from the others. However, if we are concerned with minimizing the uncertainty in the solution then the MF theory may be more robust and reliable. Moreover for the limit as $\beta \to \infty$ the MF solution becomes the minimum of the energy.

We introduce MF theory by using it to solve the problem of constructing Winner Take All (WTA) systems. This can be posed as: given a set of N inputs, T_i, to a system, how does one choose the maximum input and suppress the others? For simplicity we assume all of the T_i to be positive. We introduce the binary variables V_i as decision functions, $V_w = 1$, $V_i = 0 \;\; i \neq w$ selects element w as the winner.

We will calculate the partition function Z in two separate ways for comparison. The first uses a technique called the mean field approximation and gives an approximate answer. The second method is exact and is best for this application but cannot be used for all problems. It involves calculating the partition function for a subset of the possible V_i, a subset chosen to ensure that only one V_i is non-zero.

3.3.1 W.T.A. WITH THE MEAN FIELD APPROXIMATION

Define the energy function

$$E_1^{WTA}[V_i] = \sum_{i=0}^{N-1}\Big(\sum_{j\neq i,=0}^{N-1} V_i V_j\Big) - \lambda \sum_{i=0}^{N-1} T_i V_i \qquad (3.17)$$

where λ is a parameter to be specified. The solution of the W.T.A. will have all the V_i to be zero except for the one corresponding to the maximum T_i. This constraint is imposed implicitly by the first

term on the right hand side of (3.17) (note that the constraint is encouraged rather than explicitly enforced).

Now we formulate the problem statistically. The energy function above defines the probability of a solution for V_i to be $P = \frac{1}{Z}e^{-\beta E(\{V_i\})}$, where β is the inverse of the temperature parameter. The partition function is

$$Z = \sum_{\{V_i=0,1\}} e^{-\beta E_1^{WtA}[V_i]} \tag{3.18}$$

Observe that the mean values $\bar{V}_i$ of the V_i can be found (substituting (3.17) into (3.18)) from the identity (from which the analogy to generating functions is drawn, similar expressions can be written down for the variance of the V_i's and so on for higher order moments)

$$\bar{V}_j = \sum P(V_i)V_j = \frac{1}{\beta}\frac{\partial \log Z}{\partial(\lambda T_i)}$$

We now introduce the mean field approximation by using it to compute the partition function Z

$$Z \approx \prod_i (1 + e^{-\beta(\sum_{j\neq i,=0}^{N-1} \bar{V}_j - \lambda T_i)})$$

When calculating the contribution to Z from a specific element V_i the mean field approximation replaces the values of the other elements V_j by their mean values $\bar{V}_j$. This assumes that only low correlations between elements are important [119].

From Z we compute the mean value $\bar{V}_i = \frac{-1}{\beta}\frac{\partial \log Z}{\partial(\lambda T_i)}$ and obtain some consistency conditions on the $\bar{V}_j$

$$\bar{V}_i = \frac{1}{1 + e^{\beta(\sum_{j\neq i,=0}^{N-1} \bar{V}_j - \lambda T_i)}} \tag{3.19}$$

For the values of $\lambda < \frac{1}{T_i^{max}}$ then $\lambda T_i < 1$ and $\beta \to \infty$ (3.19) gives the correct solution of the W.T.A. problem, $V_i = 1$ for the maximum

T_i and $V_i = 0$ otherwise. For finite values of β the solution is more general and $\bar{V}_i$ assigns a weight for each value of T_i that can be used to enhance the signal.

There are several ways to solve (3.19). One consists of defining a set of differential equations for which the fixed states are solution of (3.19). An attractive method, which has been successfully applied to other problems in vision [101] and shown to be a deterministic approximation to the Glauber [55] algorithm for obtaining thermal equilibrium is

$$\frac{d\bar{V}_i(t)}{dt} = -\bar{V}_i(t) - \frac{1}{1 + e^{\beta(\sum_{j\neq i,=0}^{N-1} \bar{V}_j(t) - \lambda T_i)}}$$

3.3.2 W.T.A. WITHOUT THE MEAN FIELD APPROXIMATION

We now impose the constraint that the V_i sum to 1 explicitly during the computation of the partition function. The first term on the right hand side of (3.19) is now unnecessary and we use an energy function

$$E_2^{WTA}[V_i] = -\sum_{i=0}^{N-1} T_i V_i$$

We compute Z by summing over all possible V_i under the constraint that they sum to 1. This gives

$$Z = \sum_{V_i=0,1:\sum_k V_k=1} e^{-\beta E_2^{WTA}[V_i]} = \sum_i e^{\beta T_i}$$

In this case no approximation is necessary and we obtain

$$\bar{V}_j = \frac{1}{\beta}\frac{\partial \log Z}{\partial T_j} = \frac{e^{\beta T_j}}{\sum_i e^{\beta T_i}}$$

Thus as $\beta \to \infty$ the V_i corresponding to the largest T_i will be switched on and the other V_j will be off. This method is guaranteed to converge to the correct solution.

3.3.3 AVERAGING OUT FIELDS

The examples in the two previous sections illustrate two useful techniques. First we use the mean field approximation to derive a set of consistency conditions for the mean fields and find a deterministic algorithm for finding a solution to these equations. Second we make restrictions on the set of fields over which the partition function is to be evaluated, thereby allowing us to elegantly incorporate global constraints (such as $\sum_i V_i = 1$) without adding additional terms to the energy function. We now discuss a third powerful technique by which certain fields can be averaged out of the probability distribution. This technique is closely related to a technique proposed by Lumsdaine *et al* [93] involving the computation of marginal probability distributions. These ideas are examined further in chapter 5.

This example is based on work by Geiger and Girosi [46]. Consider the Geman and Geman formulation of image segmentation using line processes given by equation (3.9). Using the Gibb's distribution we obtain a probability

$$p(f_i, l_i) = \frac{e^{-\beta E(f_i, l_i)}}{Z}$$

where the partition function Z is given by

$$Z = \sum_{f_i, l_i} \prod_i e^{-\beta\{(f_i - d_i)^2 + (f_{i+1} - f_i)^2(1 - l_i) + l_i\}}$$

after substituting for E.

We can perform the sum over the possible states of the line process field l_i exactly as we did in the W.T.A. examples. Performing the sum over each i separately gives

$$Z = \sum_{f_i} \prod_i e^{-\beta(f_i - d_i)^2} \{e^{-\beta(f_{i+1} - f_i)^2} + e^{\beta}\}$$

We can rewrite this as

$$Z = \sum_i f_i e^{-\beta E_{eff}(f_i)}$$

where the effective energy $E_{eff}(f_i)$ is given by

$$E_{eff}(f_i) = (f_i - d_i)^2 - \frac{1}{\beta} \log\{e^{-\beta(f_{i+1} - f_i)^2} + e^{-\beta}\}$$

Thus the line process fields have been averaged out and the problem can be formulated in terms of the fields f_i only. Geiger and Girosi [46] show that in a certain limit of the parameters this becomes exactly equivalent to the formulation by Blake [14] in terms of weak constraints.

3.4 THE FORM OF THE SMOOTHNESS CONSTRAINT

We now discuss some properties that the smoothness operator S might need to have. This section reviews the approach described in [165]. We can rewrite the energy as

$$E(d(x), V_{i_L a_R}) = \int_{\cdot} \sum_{i_L, a_R} V_{i_L a_R}(d(x) - (x_{a_R} - x_{i_L}))^2 \delta(x - x_{i_L}) dx$$
$$+\lambda\{(\sum_{a,i} V_{i_L a_R} - N_l)^2 + \sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2\}$$
$$+\gamma \int_M (Sd)^2 dx$$

The Euler-Lagrange equations with respect to $d(x)$ give

$$S^2 d(x) = \sum_{i,a} V_{i_L a_R}(d(x_{i_L}) - (x_{a_R} - x_{i_L}))\delta(x - x_{i_L})$$

Here S^2 is the square of the operator corresponding to the smoothness term (i.e. $S^2 = \nabla^2$ if $S = \nabla$). There are a number of constraints for this term: (i) it must give rise to an interaction between

points that falls off at large distances,[2] (ii) it must impose enough smoothness for a minimum of the energy function to exist and (iii) it must not be too smooth, otherwise it is impossible to break the constraints with line processors.

It can be shown [165] that the interaction between the points falls off like the Green's function of the operator S^2. For example, if the operator is chosen to be the standard gradient operator, $S = \nabla$, then the Green's function behaves like $log(r)$ and blows up at infinity. A way to prevent this blow up is to incorporate cost terms for the disparity in the energy function, for example

$$\int_{M-C} (Ld)^2 dx = \int_{M-C} \left(d(x)^2 + \frac{dd(x)}{dx} \cdot \frac{dd(x)}{dx} \right) dx$$

The Green's function now obeys

$$\frac{d^2G}{dx^2} - G = \delta(x)$$

and can be shown to fall off exponentially as $x \rightarrow \infty$. The choice of such an operator satisfies condition (i).

The smoothness operator must be smooth enough to enable a minimum to exist. The amount of smoothness required depends on the dimensionality of the problem and the dimensionality of the data [40]. It is impossible, for example, to fit a membrane to isolated data points in two dimensions.

An advantage of using no higher derivatives than the gradient is that line processes are sufficient to break the constraints, and thus satisfies point (iii).

This effect of the interaction falling off as the Green's function of the square of the smoothing operator only applies for sparse data (such as edges). For dense data the precise fall off depends on the distribution of the data.

[2] This constraint is not needed for energy functions that include line processes, since the discontinuities produced by the line processes prevent interactions between points on either side of a discontinuity

3.5 ALTERNATIVE FORMS OF CONSTRAINT

So far we have considered the prior distributions as local smoothness constraints with possible discontinuities. These smoothness constraints are typically expressed in terms of Gaussian probabilities, although other probability distributions are possible without violating the convexity condition (for example see [116]). Quadratic energy functionals, which are commonly used due to their convexity, lead to Gaussian Gibb's distributions. In addition, one often assumes Markov random fields. There are, however, a number of other possible ways of expressing the priors which may be more appropriate than Gaussian distributions in sensory information processing tasks. A convenient approach is to choose a parameterized form, or set of parameterized forms, to describe the priors.

3.5.1 PARAMETRIC CONSTRAINTS

Priors or constraints that are modeled parametrically assume that the solution to an information processing problem can be expressed in terms of a parameterized form, where the choice of this form corresponds to the prior knowledge. That is, instead of putting *a priori* probabilities on the solution values we put *a priori* probabilities on the form of the solution. Technically this involves choosing the *a priori* probability $P(f)$ to be zero except on a subspace corresponding to the parameterized form. The problem then reduces to determining the values of the parameters that best fit the data (and we may solve this problem in a Bayesian fashion as well, if we have some idea of the most probable values of the parameters).

The ideas behind parametric constraints can be illustrated by the Basis Function approach to explaining the phenomenon of colour constancy in human vision [95, 164]. Although the light received from an object depends on the reflectance function of the object and

its illumination, the perceived colour of the object is relatively independent of the illumination. The human photo-receptors measure

$$l_\mu(x) = \int a_\mu(\lambda)E(x,\lambda)S(x,\lambda)\,d\lambda \tag{3.20}$$

where $a_\mu(\lambda)$ is the absorption function of the receptors ($\mu = 1$ *to* L, $L = 4$ if we consider the rods and cones), $E(x,\lambda)$ is the illumination incident on a surface with reflectance function $S(x,\lambda)$ with λ the wavelength.

The aim of theories of colour constancy is to determine the reflectance function $S(x,\lambda)$ from the measurements $l_\mu(x)$ with the illumination $E(x,\lambda)$ unknown.

This problem is clearly undetermined since, for each x, we only have a finite set of measurements to determine an infinite number of unknowns. The Basis Function approach assumes that both the reflectance function and the illuminant can be expressed as a linear combination of a set of known basis functions. More precisely

$$E(x,\lambda) = \sum_{i=1}^{N} \alpha_i(x)B_i(\lambda) \tag{3.21}$$

$$S(x,\lambda) = \sum_{j=1}^{M} \beta_j(x)H_j(\lambda) \tag{3.22}$$

where the α's and β's are unknown coefficients of the known basis functions $\{B_i(\lambda)\}$ and $\{H_j(\lambda)\}$. Experimental data, reviewed in [95], suggests that this is a good assumption for a large range of natural reflectance functions and lighting conditions with $N = M = 3$.

Substituting from (3.21) and (3.22) into (3.20) gives

$$l_\mu(x) = \sum_{i,j=1}^{i=N,j=M} \alpha_i(x)\beta_j(x)T_{ij}^\mu$$

where

$$T_{ij}^\mu = \int a_\mu(\lambda)B_i(\lambda)H_j(\lambda)d\lambda$$

The Basis Function assumption gives us L equations for $N+M-1$ unknowns at each point x (the -1 comes from the overall scaling freedom $E \to \nu E$: $S \to (1/\nu)S$, which cannot be eliminated). In general this is not sufficient to give us a unique solution and we must make further assumptions about the spatial variations of the $\{\alpha_i(x)\}$ and $\{\beta_j(x)\}$. For Mondrian worlds [84] consisting of regions of constant colour with sharp boundaries it is natural to assume that the $\{\beta_j(x)\}$ are piecewise constant (i.e. the reflectance is constant within each region) and the $\{\alpha_i(x)\}$ vary slowly with x. In particular we assume that the $\{\alpha_i(x)\}$ are locally constant across the boundaries. It can then be shown [95, 164] that, for suitable choices of N, M and L, there are a sufficient number of equations at the boundaries to solve for the $\{\alpha_i(x)\}$ and $\{\beta_j(x)\}$.

In [164] is discussed alternative assumptions on the spatial variations of the $\{\alpha_i(x)\}$ and $\{\beta_j(x)\}$ which can be used on a more general class of images. Forsythe [43] describes an alternative approach for solving for the Mondrian case.

The parametrized approach can also be used for feature recognition [168] by providing a deformable template for the feature to be extracted. In the deformable template paradigm the templates alter the value of their parameters in order to minimize an energy, which is typically some function of the parameters and image values. In a typical application, such as the face recognition system described in [168] this energy function will contain terms that act to attract the template to salient features, such as peaks and valleys in the image intensity, and image edges. The minimum of the energy function (with a given set of parameters) corresponds to the best fit of the template (as described by the parameter values) with the image. The parameters of the template are updated by steepest descent, in an effort towards minimizing the energy. This process corresponds to following a path in parameter space, and contrasts with traditional methods of template matching which would involve sampling the parameter space to find the best match. Changing the parameters corresponds to altering the position, orientation, size, and other properties of the template.

An example of such a parametrized template, and its related energy function is the following (taken from [168]). This is an "eye" template, used to locate the region in an image of a human face corresponding to an eye. The template consists of the following features:

(1) A circle of radius r, centered on a point $\vec{x}_c$. This corresponds to the boundary between the iris and the whites of the eye and is attracted to edges in the image intensity. The interior of the circle is attracted to valleys, or low values, in the image intensity.

(2) A bounding contour of the eye attracted to edges. This contour is modeled by two parabolic sections representing the upper and lower parts of the boundary. It has a center $\vec{x}_e$, width $2b$, maximum height a of the boundary above the center, maximum height c of the boundary below the center, and an angle of orientation θ.

(3) Two points, corresponding to the centers of the whites of the eyes, which are attracted to peaks in the image intensity. These points are labeled by $\vec{x}_e + p_1(\cos\theta, \sin\theta)$ and $\vec{x}_e + p_2(\cos\theta, \sin\theta)$, where $p_1 \geq 0$ and $p_2 \leq 0$. The point $\vec{x}_e$ lies at the center of the eye and θ corresponds to the orientation of the eye.

(4) The regions between the bounding contour and the iris also correspond to the whites of the eyes. They will be attracted to large values in the image intensity.

These components are linked together by three types of forces (which can be thought of as arising from the dynamics of the energy minimization process): (i) forces which encourage $\vec{x}_c$ and $\vec{x}_e$ to be close together, (ii) forces which make the width $2b$ of the eye roughly four times the radius r of the iris, and (iii) forces which encourage the centers of the whites of the eyes to be roughly midway from the center of the eye to the boundary.

The template has a total of eleven parameters; $\vec{x}_c$, $\vec{x}_e$, p_1, p_2, r, a, b, c and θ. All of these are allowed to vary during the matching.

To give the explicit representation for the boundary Yuille *et al* first define two unit vectors

$$\vec{e}_1 = (\cos\theta, \sin\theta)$$

$$\vec{e}_2 = (-\sin\theta, \cos\theta)$$

which change as the orientation of the eye changes. A point $\vec{x}$ in space can be represented by (x_1, x_2) where

$$\vec{x} = x_1\vec{e}_1 + x_2\vec{e}_2$$

Using these coordinates the top half of the boundary can be represented by a section of a parabola with $x_1 \epsilon [-b, b]$

$$x_2 = a - \frac{a}{b^2}x_1^2$$

Note that the maximal height, x_2, of the parabola is a and the height is zero at $x_1 = \pm b$. Similarly the lower half of the boundary is given by

$$x_2 = -c + \frac{c}{b^2}x_1^2$$

where $x_1 \epsilon [-b, b]$.

The potential energy function $E_c(\vec{x}_e, \vec{x}_c, p_1, p_2, a, b, c, r, \theta)$ for the image which will be minimized as a function of the parameters of the template is given below. This energy function not only ensures that the algorithm will converge, by acting as a Lyapunov function, but also gives a measure of the goodness of fit of the template. It is given as a combination of terms due to valley, edge, peak, image and internal potentials. More precisely,

$$E_c = E_v + E_e + E_i + E_p + E_{int}$$

where: (i) The valley potentials are given by the integral over the interior of the circle divided by the area of the circle,

$$E_v = -\frac{c_1}{Area}\int\int_{Circle-Area} \Phi_v(\vec{x})dA$$

(ii) The edge potentials are given by the integrals over the boundaries of the circle divided by its length and over the parabolae divided by their lengths,

$$E_e = -\frac{c_2}{Length}\int_{Circle-Bound}\Phi_e(\vec{x})ds - \frac{c_3}{Length}\int_{Para-Bound}\Phi_e(\vec{x})ds$$

(iii) The image potentials have contributions which attempt to minimize the total brightness inside the circle divided by its area,

$$E_i = \frac{c_4}{Area}\int\int_{Circle-Area}\Phi_i(\vec{x})dA$$

and maximize it between the circle and the parabolae (again divided by the area),

$$E_i = -\frac{c_5}{Area}\int\int_{Whites}\Phi_i(\vec{x})dA$$

(iv) The peak potentials, evaluated at the two peak points, are given by

$$E_p = c_6\{\Phi_p(\vec{x}_e + p_1\vec{e}_1) + \Phi_p(\vec{x}_e + p_2\vec{e}_1)\}$$

(v) The internal potentials are given by

$$E_{int} = \frac{k_1}{2}(\vec{x}_e - \vec{x}_c)^2 + \frac{k_2}{2}(p_1 - \frac{1}{2}\{r+b\})^2 + \frac{k_2}{2}(p_2 + \frac{1}{2}\{r+b\})^2 + \frac{k_3}{2}(b-2r)^2$$

The Φ's $(\Phi_v, \Phi_e, \Phi_i, \Phi_p)$ are feature maps derived from the image, and represent the data vector $\vec{d}$, in the Bayesian interpretation, from which we want to obtain the parameter vector.

The $\{c_i\}$ and $\{k_i\}$ are usually fixed coefficients but we can allow them to change values (corresponding to different epochs) as the process proceeds. Changing the values of these coefficients enable us to use a matching strategy in which different parts of the template guide the matching at different stages. For example, the valley in the image intensity corresponding to the iris is very salient and is more effective at "attracting" the template from long distances than any other feature. Thus its strength, which is proportional to c_1, should initially be large. Orienting the template correctly is usually best performed by the peak terms, thus c_6 should be large in the middle period. The constants c_2 and c_3 can then be increased to help find the edges. Finally, the terms involving the image intensity can be used to make fine scale corrections. This corresponds to a strategy in which the position of the eye is mainly found by the valley force, the orientation by the peak force, and the fine scale detail by the edge and intensity forces. In this scenario the values of the c's will be changed dynamically. Changing the weights of the different energy terms in the template matching process amounts to altering the relative contributions of multiple sensory modules in what will be referred to in the next chapter as a "weakly coupled" data fusion process.

The individual energy terms can be written as functions of the parameter values. For example, the sum over the boundary can be expressed as an integral function of $\vec{x}_e$, a, b, c and θ by

$$\int_{Para-Bound}\Phi_e(\vec{x})ds =$$
$$\frac{c_3}{Length}\int_{x_1=-b}^{x_2=b}\Phi_e(\vec{x}_e + x_1\vec{e}_1 + \{a - \frac{a}{b^2}x_1^2\}\vec{e}_2)ds$$
$$+\frac{c_3}{Length}\int_{x_1=-b}^{x_2=b}\Phi_e(\vec{x}_e + x_1\vec{e}_1 - \{c - \frac{c}{b^2}x_1^2\}\vec{e}_2)ds$$

where s corresponds to the arc length of the curve and *Length* to its total length. Note that scale independence is achieved by dividing line integrals by their total length and double integrals (over regions) by their area.

The minimization is done by gradient descent on the energy function in parameter space. The update rule for a parameter, for example r, is given by

$$\frac{dr}{dt} = -\frac{\partial E_C}{\partial r}$$

If we look at this deformable template example from the Bayesian viewpoint we see that the *a priori* information is used to specify that the solutions are to be members of a certain class of parametrized functions. All other solutions are excluded (i.e. have zero *a priori*

probability). Within this parametrized class all functions are *a priori* equally likely. We have not provided any *a priori* bias for solutions within the parametrized set of functions. An obvious extension of our approach would be to put a nonuniform *a priori* distribution on the functions within the parametrized class. The image formation model is characterized by the energy E_c. It is this energy that allows us to choose between elements of the parametrized function space based on the image data.

Often one must specify an algorithm with a very general parametrization of the solution, since little may be known about the actual form of the solution. In such a case, there may be far too many parameters in the model to practically determine the solution. The only recourse that one is usually left with is to make the parametrization more specific. In doing so the algorithm designer runs the risk of having the algorithm fail if the parametrized model does not adequately represent the solution. In other words, the *a priori* constraints built into the parametrized model may be invalid if the constraints are not general enough, yet if the model is too general the task of finding a solution is not significantly aided by the assumption of the model. This dilemma is considered further in chapter 4, where we introduce the concept of constraint adaption in strongly coupled data fusion, wherein the *form* of a parametrized image formation or prior model, and not just the parameter values, are adapted by using inter-module consistency measures. In chapter 4 we will present a similar approach that uses temporal consistency measures (either within a single module or between many modules) to adapt the image formation or *a priori* models.

3.5.2 MINIMAL DESCRIPTION LENGTH CODING

The Minimal Description Length (MDL) approach developed by Risannen [133] for statistical applications suggests that we should interpret the data in terms of the minimal description (subject to some error bounds) with respect to a previously specified model. In

this method, which provides an elegant connection between information theory and probability theory, the length of the description can be related to a probability (so that solutions with longer descriptions are less probable). Risannen shows that the MDL criterion can be formally related to the Bayesian approach, with the *a priori* probabilities $P(f)$ corresponding to the length of the description and the $P(d|f)$ term being related to the error in the description.

The crucial issue in this approach is how the model should be defined and coded. While Risannen has some results relating this to optimal coding theory it is unlikely that they are relevant to sensory information processing. This is due to the observation that the optimal coding theory presupposes a discrete system, which is unlikely to found in natural systems such as the brain. The choice of the model in the MDL approach must correspond to the constraints being imposed. This leads to theories similar to those described before, and some examples will be given shortly. But the MDL approach provides an interesting perspective, potentially useful for some novel applications, and is philosophically very attractive. The idea of describing an image scene in terms of the minimal description with respect to a specific model has a strong intuitive appeal.

The parametrized constraint approach, described in the previous section, can be directly interpreted in terms of MDL. The choice of parameters corresponds to the best description of the data in terms of the parametrized model. This is a fairly trivial example of MDL since we are only minimizing the error of the model to the data and we are not altering the length of the description (unless we made the description of some parameter values "longer", and hence less probable, than others). Observe that for the Basis Function approach to Colour Constancy the choice of the model, i.e. the Basis functions, is determined by analysis of the real world. For example, one can perform a principal component analysis on a large set of colour images in an effort to obtain a small set of Basis functions which characterize the observed data (i.e. by finding the eigenfunctions of the Karhunen-Loeve matrix of the measurements with the largest eigenvalues [86]).

A deeper example of MDL can be found in the work of Leclerc [85] (see also the related work of Keeler [78]) on using MDL methods in performing image segmentation tasks. Leclerc models the image segmentation process as one of finding the minimal length description in terms of piecewise constant surfaces (he also discusses the generalization to polynomial surfaces) and determines basic costs for the descriptions of the surfaces and their boundaries. This approach leads to a method which has strong similarities to the Markov Random Field approaches to image segmentation (Geman and Geman, et al [51]) when the smoothness coefficient is made very large [171].

3.5.3 MULTIPLE SETS OF PRIORS

MDL suggests a generalization of the parameterized constraint approach by considering a model containing several possible parameterized surfaces. The data is described by the minimal number of surfaces with optimal parameter values. Recent work described in [171] applies MDL ideas to problems involving stereo and motion transparency. The idea in this approach is to represent the data in terms of the minimal number of parametrized surfaces that best match the data. Another form of this approach would involve a set of (possibly quite different in form) different parametrized models. One would then apply the model for which the data results in a solution with the minimum description length (with respect to the model). For example one could have the following parametrized models for surfaces: (1) Planar Surfaces, (2) Piecewise Quadric Surfaces, (3) Piecewise Sinusoidal Patches, and (4) Harmonic Surfaces. Each of these models may have only a few parameters, so that the minimization process is not too severe. The models are different enough, however, that, for a given data set, one of them will be the most suitable. One can determine the suitability of a model, if the model is expressed in MDL terms, by looking at the length of the minimum description length solution to the measured data provided by the application of the model.

The methods described above are essentially data fusion methods, wherein the information being fused is the *a priori* information, rather than the measured data. Recall the discussion in section 2.6 on multisubjective priors in Bayesian methods. The type of data fusion that is performed here (it might more correctly be termed "information fusion") is of a different sort than the data fusion operations most commonly encountered in the sensory information processing literature. In the standard approaches to data fusion, one has a number of pieces of data or measurements and one combines them, subject to some *a priori* information to obtain the result most consistent with the measurements and the *a priori* constraint. In the next chapter we will provide a more complete characterization of such methods, and will refer to them as *weakly coupled* data fusion algorithms. The data fusion methods that we have been discussing in this chapter, however, involve taking a single measurement (vector) and a set of *a priori* constraints (multisubjective priors) and combining them in a way that produces the result that is most consistent with the data. The actual prior to be used is either one of the priors (winner take all prior selection) or some modified version (fused, or consensus) of the priors. In the next chapter we will refer to such methods as a form of *strongly coupled* data fusion.

The distinction between these two types of data fusion is an important one and should be emphasized. The first approach combines multiple sets of data using a single *a priori* constraint, while the second approach combines multiple *a priori* constraints using a single set of measurements. Much work has been done on the first type of data fusion, but little has been done, in the sensory information processing research community, on developing systems based on the second form of data fusion. There is, however, evidence that biological systems frequently utilize such fusional techniques, wherein the *a priori* constraints used in solving a sensory processing problem are selected from a set of possibilities.

3.6 CHAPTER SUMMARY

- Smooth energy function minimization is a special case of the Bayesian formulation which is well suited for imposing smoothness constraints. A suitable choice of the smoothness operator makes the solutions linear combinations of basis functions.
- By defining a Gibb's distribution one can give a probabilistic interpretation to minimizing energy functions. This shows that energy functions are a special case of the Bayesian formalism corresponding to Markov Random Field distributions.
- Markov random fields can be implemented with binary "line process" fields, which results in a special case of the Bayesian approach that allows the "breaking" of smoothness constraints at places where the smoothness constraint is inappropriate.
- We can generalize Markov random fields by adding in binary valued "Matching Element" fields. These matching fields allow us to specify correspondences between features in separate images, such as is required in long-range motion analysis or stereopsis.
- For many vision problems, in particular those involving correspondence, there are global constraints on the field which need to be satisfied. Statistical techniques give ways of imposing these constraints absolutely, unlike the more traditional weak methods of adding terms to the energy function which merely "encourage" the constraints to be satisfied.
- The use of parametric constraints results in a special case of the Bayesian formalism which is an alternative to the standard smooth surface with/without discontinuities *a priori* constraint. It assumes a parameterized form for the solution and determines the parameters to best match the data.
- Another special case of the Bayesian formalism can be specified which interprets the data in terms of the minimal description with respect to a prior model and minimal error criterion. The prior model could include a set of parameterized models, hence this method would determine which parameterized form(s) should be used for a specific set of data.
- A form of data fusion results when one uses a multiplicity of *a priori* constraints, either by selecting one out of a set of possibilities, or by finding a consensus between the *a priori* opinions, in order to extract information from a single set of sensory measurements.

Chapter 4

Weakly vs. Strongly Coupled Data Fusion: A Classification of Fusional Methods

Chapter 2 introduced the Bayesian approach to the processing of sensory information. One of the notable aspects of the Bayesian formulation is the ease with which constraints can be embedded. The constraint embedding was seen to be performed through the specification of suitable image formation and prior models. The image formation model typically involved the physical, and to a lesser extent natural, constraints, while the prior model typically involved natural and artificial constraints.

We saw in chapter 1 that it is often desirable to fuse information from many sensory information processing modules to solve a given problem. One can think of this data fusion process as one of embedding constraints, where the constraints are the pieces of information from some subset of the sensory modules. In the Bayesian

formulation of data fusion, then, these sources of information would be used to specify, or alter, the image formation and prior models, of some Bayesian estimation process. This line of thought leads to a classification of data fusion algorithms, wherein we distinguish between different types of fusion algorithms by the ways in which the constraints (i.e. information from the various sensory modules) are embedded in a particular sensory information processing task.

We propose in this text just such a general classification of data fusion algorithms, and describe it in detail in the next section.

4.1 A CLASSIFICATION OF FUSIONAL METHODS

In this section we introduce our classification of data fusion algorithms. This classification is based on the consideration of data fusion as the embedding of constraints. Our classification involves two major classes of fusional methods, that of *weakly* versus *strongly* coupled data fusion. In weak coupling the outputs of two or more algorithms that produce information independently are combined. From the Bayesian viewpoint, in a weakly coupled fusional algorithm the image formation and prior models do not depend on any of the sensory modules, and the outputs of the sensory modules are considered purely as data. Most of the current methods of sensory fusion use weakly coupled approaches.

In a strongly coupled fusion algorithm, on the other hand, the operation of one sensory module is affected by the output of another sensory module, so that the outputs of the two modules are no longer independent. From the Bayesian point of view, in a strongly coupled fusional algorithm, the outputs of one, or more, of the modules are used to specify or alter the image formation and/or prior models of some or all of the sensory modules (including, possibly itself).

There are useful distinctions to be made within the broad classes of weakly and strongly coupled fusion algorithms. These deal with the specific details of how the pieces of information are actually combined. These more detailed classifications are described in the next two sections, along with more in depth descriptions of the ideas behind weak and strongly coupling.

4.2 WEAKLY COUPLED DATA FUSION

As was mentioned in the introductory chapter, there are three main reasons for performing sensory fusion. To emphasize the distinction between fusional methods which have different goals in our discussions, we propose a partitioning of the class of weakly coupled fusional algorithms into three distinct subclasses. Each of these represent a specific method for weakly fusing sensory data. This partitioning is illustrated in figure 4.1.

4.2.1 CLASS I WEAKLY COUPLED DATA FUSION

The subclasses shown in figure 4.1 represent a natural division of approaches to sensory fusion. The first, and weakest, sort of sensory fusion that is treated in our categorization is a weighted combination of the outputs of two or more sensory modules. In this class of fusion the sensory modules whose outputs are to be combined must be stable and provide unique solutions. Thus, the only purpose of the fusion in this case is to reduce the uncertainty of the resulting fused measurement. The weighting of one information source with respect to another is derived from measures of the relative reliabilities of the two information sources. These reliabilities must be determined during the operation of the sensory modules. If reliability measures are not available one must, in the absence of any other information, assume that each source of information is equally reliable.

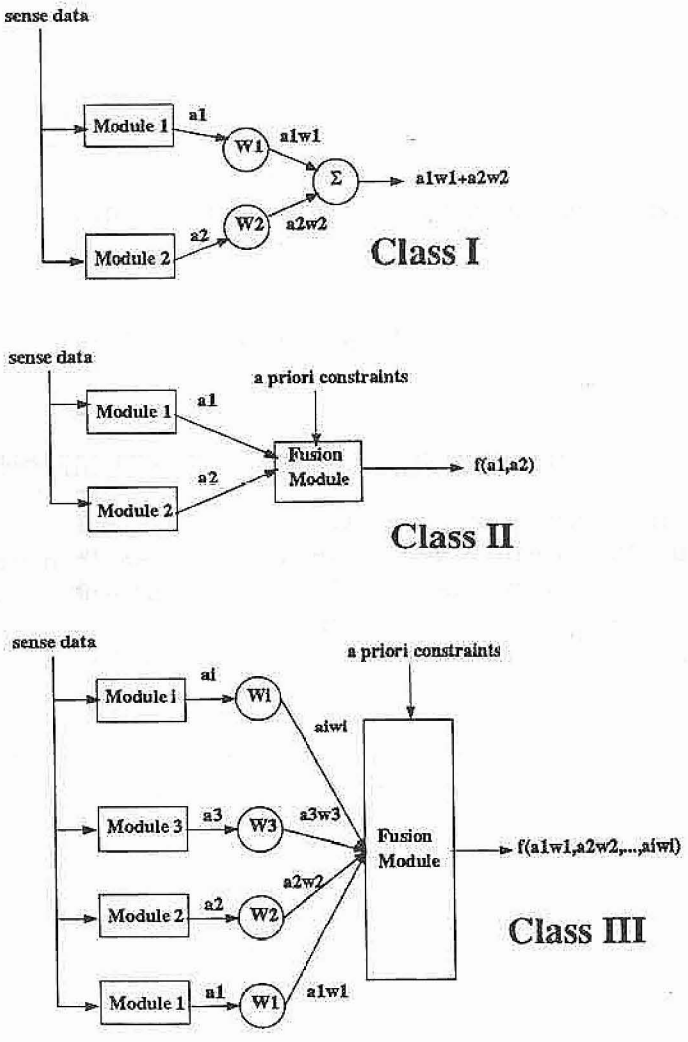


Figure 4.1: A utility based classification of weakly coupled fusional algorithms.

As an example of a class I weakly coupled fusion algorithm consider the following energy function:

$$E(\vec{f}) = \int \alpha_1(\vec{d}_1 - \vec{D}_1(\vec{f}))^2 + \alpha_2(\vec{d}_2 - \vec{D}_2(\vec{f}))^2 d\vec{x}$$

In the above example, we determine our desired information $\vec{f}$ as that which minimizes $E(\vec{f})$ given the two sources of data $\vec{d}_1$ and $\vec{d}_2$. The functions D_1 and D_2 are the image formation models for our two modules. These represent the process by which a given solution vector $\vec{f}$ gives rise to the data, $\vec{d}_1$ and $\vec{d}_2$. The constant factors α_1 and α_2 are typically the relative reliabilities of the two sources of data. Suppose we assume that f and the d_i are scalar, and that the functions D_i are particularly simple, $D_i(f) = f$. Then the f that minimizes E is seen to be

$$f_{optimal} = \left(\frac{\alpha_1}{\alpha_1 + \alpha_2}\right) d_1 + \left(\frac{\alpha_2}{\alpha_1 + \alpha_2}\right) d_2$$

In this simple case, then, the data fusion corresponds to simple weighted averaging, where the weights are related to the reliability of the information.

4.2.2 CLASS II WEAKLY COUPLED DATA FUSION

The second class of fusional algorithms combine two or more sources of information in order to yield a unique, robust, solution. It is assumed that, in this case, each of the sensory modules taken alone can not provide a unique solution. In this class of fusional algorithms the relative reliabilities of the information sources are irrelevant since all of the information sources are required for a unique solution. This class of fusional algorithms includes the approaches collectively referred to as active vision in [2, 1]. These methods are often posed in an algebraic setting, where exact results can be obtained, although this is not always possible. A simple, illustrative, example of an algebraic fusion method is the following. Suppose we wanted to determine the resistance of a resistor, and we had

available to us a battery of unknown voltage, a voltmeter and an ammeter. We could measure the voltage of the battery with the voltmeter giving us V_b. We could then connect the resistor to the battery in series with the ammeter. The ammeter would then give the current flowing in the resistor, I_b, due to the EMF of the battery. The resistance, R can be calculated from V_b and I_b through the equation $R = V_b/I_b$. Thus we have fused the data values V_b and I_b to determine the desired quantity. Note that we could not have determined R without knowing both V_b and I_b unless we imposed some *a priori* constraint regarding the value of the voltage or current (i.e. assuming V_b to be some value known *a priori)*. The "image formation model" in this case is Ohm's law, $I = V/R$.

One significant aspect of class II fusional algorithms is that they are prone to be excessively sensitive to noise in the data. Since the data values are absolutely required there is no alternative but to use the data, even if it is very noisy. This drawback of algebraic fusion is illustrated quite convincingly in chapter 7, and is present in the examples of class II fusion we describe below, even if no mention of these difficulties are given in the original references.

4.2.3 CLASS III WEAKLY COUPLED DATA FUSION

The third class of fusional algorithms can be thought of as a combination of the methods in the first two classes. In this class of sensory fusion methods the component sensory modules are inadequate in that they cannot provide a unique solution. However, the additional information required to obtain a unique and stable solution is not provided solely by other sensory modules. This information comes in part from prior constraints added in the fusional process. Adding in these constraints allows one to weight the contributions of the outputs of the sensory modules, as well as the reliance on the assumed constraints, since not all of the information sources are required to get a unique solution.

Class III weakly coupled data fusion is useful in reducing the uncertainty resulting from noisy data. For example, in the resistance measurement process described earlier, the voltage and current measurements may be noisy, resulting in uncertain resistance values. If the noise in the voltage and current measurements at different times are independent the uncertainty in the derived resistance values can be minimized by averaging the voltage and current values obtained over a number of measurements taken at different times. In this case we would have that $R = \sum_i V_b(t_i)/\sum_i I_b(t_i)$. Note that, in this form of fusion, not all of the data is absolutely necessary in order for a value of R to be obtained. We can compute a value for R using only one measurement of V_b and I_b. The additional measurements are used only to reduce the effect of noise.

In the class III approach one has the extra freedom of weighting the contribution the various measurements have toward the computation of the desired parameter. If a measurement is known to be particularly noisy it can be weighted by a small factor, while if the noise on a measurement is known to be low, then that measurement can be weighted highly. In the example of the resistance measurement we could compute R as a ratio of weighted sums:

$$R = \frac{\sum_i \alpha_i V(t_i)}{\sum_i \beta_i I(t_i)} \frac{\sum_i \beta_i}{\sum_i \alpha_i}$$

Note that we cannot weight the voltage measurements versus the current measurements; we can only weight voltages measurements relative to other voltage measurements, and current measurements relative to other current measurements. This reflects the fact that the voltage and current measurements must be independent in order to compute the resistance. Both are required. This is an illustration of the difference between the class III fusional methods and the class I fusional methods.

4.3 STRONGLY COUPLED DATA FUSION ALGORITHMS

The other set of fusional algorithms contained in our categorization are those corresponding to the strongly coupled fusional algorithms. The main feature of the strongly coupled approaches to sensory fusion, and which distinguishes these methods from weak methods, is that the operation of one or more modules can be affected by the results of other modules. In other words the outputs of the component modules are no longer independent. A simple example is a feature matching stereo vision algorithm where the matching of features is guided by depth values available from other sources (typically these sources are the outputs of other sensory modules, information in the form of *a priori* constraints can be used but we consider such a method to be weakly coupled, class II, as these additional constraints can be thought to be a part of the stereo module itself).

The development of strongly coupled data fusion algorithms are motivated by the idea, expressed in chapter 1, that data fusion should be concerned with reducing dependence on possibly invalid prior constraints, and not just on reducing the level of uncertainty in the value of a parameter.

As in the case of weakly coupled data fusion algorithms, we split strongly coupled algorithms into a number of different classes. This subdivision is based on the ways in which the modules to be fused interact with each other. We classify strongly coupled data fusion algorithms into the following basic types:

- feedforward prior constraint adaption.
- feedforward image formation model adaption.
- recurrent constraint and image formation model adaption.

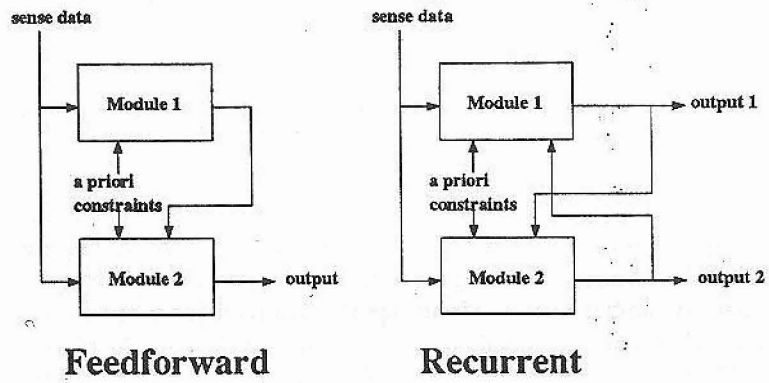


Figure 4.2: A categorization of algorithms for strongly coupled data fusion.

The structure of these types of strongly coupled data fusion algorithms are depicted in figure 4.2.

The feedforward prior constraint adaption operates by having the prior assumptions or constraints that a module operates under change due to information from one or more otherwise independently functioning modules. The feedforward image formation model adaption operates by having the model that a module uses to represent the process of image formation change due to information from one or more otherwise independently functioning modules. Recurrent strong coupling occurs when a module that is being affected by another (through adaption of constraints or image formation model) subsequently affects the other module. There are four basic types of recurrent strong coupling for two module interaction, as indicated by the diagram. If more modules are used more combinations of recurrent and feedforward coupling are possible.

The distinctions between weakly and strongly coupled fusional processes may become blurred if we allow complicated interactions involving feedback, but they are a useful way to characterize existing

systems. If the output from the two modules is independent then weak coupling methods are appropriate, otherwise strong coupling should be used.

In the following subsections some simple illustrative examples of the different types of strongly coupled approaches to fusion are given.

4.3.1 STRONG COUPLING BY PRIOR CONSTRAINT ADAPTION

In prior constraint or model adaption, fusion occurs by one, or more, modules changing the *a priori* constraints on another module (or modules). A simple example would be where a smoothness constraint is altered, such as in a stereo vision algorithm. One could use the output of an occluding edge (where depth values are usually discontinuous) detection algorithm to "break" the smoothness constraint at occluding edges, thereby keeping the depth extraction module from smoothing over the depth discontinuities. This process can be described as an energy minimization process where the energy to be minimized is the following:

$$E(\vec{x}) = \int (I_l(\vec{x}) + I_r(\vec{x} + \vec{D}(\vec{x}))^2 + (e(\vec{x}) - 1)||\nabla \vec{D}(\vec{x})||^2 d\vec{x}$$

where $\vec{D}(\vec{x})$ is the disparity field, I_l and I_r are the left and right image intensity fields, and $e(\vec{x})$ is the (binary) field of occluding edge locations ($e(\vec{x}) = 1$ at an occluding edge, and is zero elsewhere). The occluding edge field is produced from a module that is independent of the disparity field estimation process.

4.3.2 STRONG COUPLING BY ADAPTION OF THE IMAGE FORMATION MODEL

In image formation model adaption, fusion proceeds through one or more modules specifying or altering the image formation model

used by a particular information processing module. Typically the image formation module depends on a number of parameters which must either be assumed or determined by independent modules. For example, if a Gaussian image formation (or sensor noise) model is assumed, then the variance and mean of this Gaussian must be known. If some independent module determines these parameters then the Bayesian estimation process is strongly coupled to that module. In vision applications the image formation model is generally more complicated, but still can be parametrized. For example, in a shape from shading module, a Lambertian reflectance function may be assumed, with the light source intensity and direction, and the surface albedo as parameters. If these parameters are not assumed, but are estimated by some other, independent, modules then the shape from shading process would be strongly coupled to these modules. An example of this type of strongly coupled data fusion applied to shape from shading is detailed in chapter 7. In that application, the following energy function is minimized to produce the solution to the shape from shading problem.

$$\int [(E(\vec{x}) - w_l(\vec{x})(\hat{n} \cdot \hat{s}) - (1 - w_l(\vec{x}))(\hat{k} \cdot \hat{h})^m)^2 + \mu(||\hat{n}|| - 1)^2 + \lambda(||\nabla \hat{n}||^2)] dA + \int_\Gamma C dl$$

Here $E(\vec{x})$ is the data array. The image formation model is represented by the $(E(\vec{x}) - w_l(\vec{x})(\hat{n} \cdot \hat{s}) - (1 - w_l(\vec{x}))(\hat{k} \cdot \hat{h})^m)^2$ energy term. The array $w_l(\vec{x})$ comes from some independent sensory module and represents a segmentation of the image E into regions of primarily specular or primarily Lambertian reflectance. Thus the image formation model is altered by the segmentation module, and hence the shape from shading process is a strongly coupled one.

4.3.3 RECURRENT STRONG COUPLING

Our final class of strongly coupled fusion algorithms involves those in which either the prior or image formation model used by a module can be affected by the output of a module which is itself

affected in some way by the original module. In this way a feedback loop is created, causing recurrent behavior. It is clear that one must worry about convergence and stability in these cases as the resulting system is a dynamic one, and may diverge or oscillate, or even show chaotic behavior.

As an example of a recurrent strongly coupled data fusion process, consider the shape from shading example described in the previous section. If the segmentation module requires some knowledge of the object shape in order to produce a segmentation, and if this shape information comes from the shape from shading module which depends on the segmentation, then the shape from shading process is clearly recurrent. This example is examined in more detail in chapter 7, where the dynamics of the recurrence are briefly considered.

4.3.4 COUPLED MRF METHODS AS STRONGLY COUPLED DATA FUSION

It can be seen that coupled Markov random field methods can be viewed as being recurrent strongly coupled data fusion methods. To choose a specific example consider the Geman and Geman [51] coupled MRF approach to image segmentation. Here the two coupled modules are image intensity estimation (or image smoothing) and discontinuity detection. This segmentation algorithm can be represented as one of minimizing the following energy function (this is the same equation as (3.9))

$$E(f_i, l_i) = \sum_i \{f_i - d_i\}^2 + \lambda \sum_i \{f_{i+1} - f_i\}^2 (1 - l_i) + \mu \sum_i l_i$$

The location of the line processes l_i (corresponding to estimates of the location of the intensity discontinuities) depend on the estimate of the image intensities, and the estimate of the image intensity field itself depends on the location of the discontinuities. The recurrent nature of the fusion of the discontinuity and intensity field modules is readily apparent.

One can construct an implementation of the two modules by the following iterative procedure. First assume that there are no discontinuities (i.e. initialize l_i to be zero for all i). With this constraint determine the set f_i that minimizes E. Taking this set of f_i values as a fixed constraint, determine the line process field $\{l_i\}$ which minimizes E. This process is then cycled, alternating between fixing l_i and solving for f_i and fixing f_i and solving for l_i. As with any recurrent fusion algorithm one must be concerned about the convergence and stability of the recurrence. It is not clear whether the above algorithm will converge to the global minimum of E (over the entire space of l_i and f_i fields) even if at each step the global optimum of E (over the entire space of l_i or f_i, fixing l_i or f_i) is found. The final solution will probably depend on the initial condition. We should use any *a priori* information we may have (which could come from an independent discontinuity detection module or segmentation module) in determining a suitable starting point for l_i (or for f_i if we start that way).

4.4 BAYESIAN IMPLEMENTATION OF DATA FUSION

Our constraint centered approach to data fusion, and the distinction between weakly and strongly coupled fusion, is most clearly illustrated by a Bayesian formulation of sensory information processing tasks.

We saw how, in the Bayesian formulation as represented by equation (2.1), *a priori* constraints can be combined with the output of a sensory module (i.e. the data $\vec{d}$). This is done by specifying the constraints in the *a priori* probability $P(\vec{f})$ and the image formation model $P(\vec{d}|\vec{f})$. However, we can extend the utility of the Bayesian approach to allow the fusion of multiple sources of data. We can

write:

$$P(\vec{f}|\vec{d}_1,\vec{d}_2) = \frac{P(\vec{d}_1,\vec{d}_2|\vec{f})P(\vec{f})}{P(\vec{d}_1,\vec{d}_2)} \quad (4.1)$$

A critical question concerns the independence (or dependence) of the two data sources. If we can assume independence then

$$P(\vec{d}_1,\vec{d}_2|\vec{f}) = P(\vec{d}_1|\vec{f})P(\vec{d}_2|\vec{f}) \quad (4.2)$$

The methods in class I weak coupling are easily put into the Bayesian framework as the solution $\vec{f}$ is obtained from finding the $\vec{f}$ which maximizes the conditional probability in equation (4.1). The prior distribution on the possible outputs, $P(\vec{f})$ is taken to be flat or uniform, since there are no *a priori* constraints on the $\vec{f}$'s.

The class III fusional methods are most easily formulated as energy function minimization problems, which can be converted to Bayesian problems through the application of the Gibb's distribution as described earlier.

Uncertainties in the input data are modeled by the conditional probabilities $P(\vec{d}_i|\vec{f})$ (i.e. the image formation model). If the data are highly certain then this distribution will be highly peaked for the optimal $\vec{f}$. If the data is uncertain the distributions will be more spread out over the space of the $\vec{f}$'s. Uncertainties in the output can be measured by the form of the distribution $P(\vec{f}|\vec{d}_1,\vec{d}_2,...)$. If this distribution is highly peaked about the output $\vec{f}$ then the uncertainty in the output is low. If, however, this distribution is flat near the output value then the output uncertainty will be high.

In energy function terminology the above Bayesian formulation corresponds to linearly summing weighted (quadratic) data consistency terms from different sources. The weights on each term are related to the sharpness of the conditional probabilities densities, $P(\vec{d}_i|\vec{f})$. The energy function approach to fusing data has been used by Poggio and coworkers[122].

When the independence assumption is valid and we can use the above approach to sensory fusion, we say that the fusional method

is *weakly* coupled, where the coupling referred to is between the information sources. The independence assumption may simplify the calculations but is certainly not always justified. It is not always easy, however, to determine the dependence between different data sources. There is also a pragmatic balance between the increased accuracy attained by using dependent, or strong, coupling and the possible additional computational cost.

Strong coupling occurs when we do not assume independence of data sources. An example of this is described in detail in the next chapter, and in chapter 5. In chapter 5 we present an attempt at trying to couple a stereo algorithm with a monocular depth algorithm, such as depth from defocus or controlled eye movement. The independence assumption would imply that, given an object, the stereo data (the projection of significant features of the object onto the two eyes) was independent of the monocular depth data. This, however, is not the case since features in the left eye with a certain monocular depth estimate are likely to correspond to features in the right eye with similar monocular depth estimates. So in this case the independence assumption is unjustified and a more sophisticated strong coupling method should be used.

If the data sources are not independent then equation (4.2) is no longer valid and we must, if the Bayesian approach is still to be used, determine the form of $P(\vec{d}_1,\vec{d}_2,...,\vec{d}_n|\vec{f})$. Often, this distribution may be difficult to specify; in these cases casting the problem in the dual energy functional framework may result in a more tractable problem, as is done in section 5, wherein a strongly coupled fusion algorithm for stereo vision is described in terms of an energy function minimization process.

A second approach to the Bayesian formulation of strongly coupled fusion is to treat the problem as a single sensory module combined with *a priori* constraints, as in the case of a weakly coupled algorithm. However, now we allow for the adjustment of the constraints by the outputs of other sensory modules. This has the effect of altering the probabilities $P(\vec{f})$ in a data dependent way. There

are two main ways in which this adjustment of the $P(\vec{f})$ distribution can be accomplished. The first is to have a set of $P(\vec{f})$ distributions, each of which is appropriate in a given domain. The other sensory modules can then be used to decide which domain is being operated in (this is the "multisubjective" approach mentioned in chapter 2). Once this decision has been made the appropriate distribution for $P(\vec{f})$ can be selected and then used to determine $\vec{f}$ given the original sensory data. Alternatively, one could provide the $P(\vec{f})$ directly from another sensory module, instead of merely selecting from a fixed set of possible distributions. In any event, we would characterize the Bayesian implementation of a general strongly coupled data fusion algorithm as follows. Such a system can be described as determining the parameter(s) f which optimize some statistic of the following conditional density:

$$p(\vec{f}|\vec{d}) = \frac{p(\vec{d}|\vec{f}; z_1, z_2, ..., z_n)p(\vec{f}; z_1, z_2, ..., z_n)}{p(\vec{d}; z_1, z_2, ..., z_n)}$$

where $\vec{d}$ is the data being input to the Bayesian parameter estimation module from some sensor or sensory module, and $(z_1, z_2, ..., z_n)$ are the data from n sensors or sensory modules (which may include the sensor producing $\vec{d}$ or the module producing $\vec{f}$). The image formation model $p(\vec{d}|\vec{f}), p(\vec{d})$ as well as the *a priori* output model $p(\vec{f})$ are all seen to be functions of the data $(z_1, z_2, ..., z_n)$.

The data (z_i) can come from strong or weak fusional modules as well. In fact there is no reason that the output z_k of a given module that is strongly coupled to the module that determines f can not itself be influenced by f. Such a system would be recurrent in its operation and care must be taken in the design of such a system to prevent unstable behavior (such as limit cycles or chaos in the values of z_k and f).

Note that this formulation includes both recurrent and feedforward adaption of both the prior and image formation models. In the recurrent case some knowledge of the system dynamics will be required in order to formulate the *a posteriori* probability correctly. Proper design of recurrent methods is still very much an open

research problem. We, and others, have only used *ad hoc* approaches in obtaining stable and convergent recurrent strong coupling (see for example the algorithm of Singh later in this section, as well as the strongly coupled temporal coherence edge detection algorithm described in chapter 8.)

In the design of strongly coupled fusion algorithms the question arises as to whether it is better to adapt the prior model or to adapt the image formation model. As pointed out by Szeliski (personal communication, also see [147]), whether you adjust the smoothness constraint (prior model) or the data constraint (image formation model) would not matter if all the probabilities were point-wise independent. But since smoothness constraints are not (they introduces some spatial correlation into the solution), weakening or strengthening them may introduce "flat spots" or other artifacts into the solution. If you have a reasonable image formation model, this is where to adjust for the variation in the quality/reliability of the data. If, however, you know something about the surface, e.g., that it breaks or creases, adjusting the smoothness constraint is appropriate.

4.5 EXAMPLES OF WEAKLY COUPLED DATA FUSION IN THE VISION LITERATURE

The literature on data fusion is vast and somewhat chaotic, and we cannot expect to provide a complete review of it here. We will instead briefly describe a relatively small number of examples of existing approaches to data fusion that are consonant with our philosophy. Later chapters will give some more detailed examples of our own application of the constraint based approach to data fusion.

A common application of data fusion is to combine enough sources of information to allow a unique solution to a given sensory in-

formation processing task to be obtained. These methods, of which three approaches are described below, are examples of class II weakly coupled data fusion algorithms.

In [3] Aloimonos and Schulman present a number of data fusion methods (they refer to the fusional process as *integration of visual modules*) that use the controlled motion of a camera to provide additional images which can be used to make ill-posed problems well-posed. They refer to these methods as *active vision.* Many of the active vision methods are examples of class II weak coupling, as they use the extra information provided by the time sequence of images to provide unique solutions to problems which are otherwise underdetermined.

A good example of the active vision approach is the following active shape from contour algorithm, taken from [3]. The goal here is to determine the slant p and tilt q of a plane on which a polygonal contour is inscribed. It is assumed that we have a binocular pair of images available, taken from cameras having parallel optical axes. From the images we measure the areas S_L and S_L of the (paraperspective or perspective) projection of the polygonal contour onto the left and right image planes. In addition we compute the centers of mass, $(A_L, B_L), (A_R, B_R)$ of the left and right images of the contours. We can specify the following constraint (derived in [3]) between the desired parameters p, q and the above computed quantities:

$$\frac{S_L}{S_R} = \frac{1 - A_L p - B_L q}{1 - A_R p - B_R q}$$

It is clear that the above constraint is not enough to uniquely define both p and q. In order to obtain a unique solution we must add more information. The active vision approach to this is to move the cameras and obtain a new set of images, which will give (for perspective projection) a new constraint that is independent of the original constraint. If the cameras are moved by rotating them (tilting them upwards for example) by a small angle θ, then the new constraint is:

$$\frac{S'_L}{S'_R} = \frac{\cos\theta + q\sin\theta - A'_L p - B'_L(q\cos\theta - \sin\theta)}{\cos\theta + q\sin\theta - A'_R p - B'_R(q\cos\theta - \sin\theta)}$$

where A'_L etc. are the measurements made on the images from the rotated cameras. We now have two constraint equations for our unknowns p and q so that a unique solution is available (a slight modification is required for the degenerate case when $p = 0$ [3]).

Another example of an algebraic, or class II, approach to data fusion is that of Photometric Stereo. This is a method, introduced by Woodham, [162] for determining the shape of an object from the information contained in two different views of the object obtained by changing the position of the source of illumination. While this is not commonly thought of as a data fusion algorithm, it can be considered as one, and, in particular, is an example of a class II weakly coupled fusion algorithm. The operation of the shape from photometric stereo algorithm is based on the observation that the image $I(\vec{x})$ capturing the observed brightness of an object (having uniform reflectance properties) depends on the direction and intensity of the illumination source $(\vec{s})$ as well as the unit surface normal vector field $\hat{n}(\vec{x})$ of the object. This is summarized by the reflectance law:

$$I(\vec{x}) = R(\hat{n}(\vec{x}), \vec{s})$$

The function R is often termed the *reflectance map,* as it represents the mapping between the object shape $(\hat{n}(\vec{x})$ and the observed image values $(I(\vec{x}))$. If we fix $\vec{s}$ to be some known value, and if we know the form of R (e.g. $R(\hat{n}, \vec{s}) = \hat{n} \cdot \vec{s}$ for a Lambertian surface), then the above equation gives a one dimensional constraint on the two dimensional (one for each component) unit surface normal vector field. Thus, we need additional information in order to obtain a unique solution for $\hat{n}(\vec{x})$. In photometric stereo this extra information is obtained by changing the direction of the light source (i.e. changing $\vec{s}$).

The photometric stereo process is similar to an active vision algorithm except that here the camera does not move, the light source does. It is clearly in the same spirit as active vision - alter the environment through the degrees of freedom available to the system in a way which serves the perceptual process.

The work of Nandhakumar and Aggarwal [114, 115] on fusing thermal and visual imagery also uses a class II, algebraic, weakly coupled approach. From independent modules they obtained measures of thermal and visual irradiance. These two measures were combined in algebraic expressions for estimating the heat fluxes at object surfaces. Their approach is similar in spirit to the active vision algorithms such as the one just described in that both the thermal and visual information is required in order to obtain unique solutions for the desired surface parameters. The fusional process is an algebraic one, and is what we have termed a class II weakly coupled fusion algorithm. The details of the algorithm can be found in [114], but we will repeat some of them here to make clear the form of the fusion that is performed.

The aim of the algorithm is to obtain the value of a parameter that can be used as a feature in image segmentation and object recognition. This parameter was the ratio $\frac{W_{cd}}{W_{abs}}$ of the heat conducted from the surface of an object to its interior, W_{cd}, to the portion of the irradiation absorbed by the surface of the object, W_{abs}. These two heat fluxes are determined from a thermal image and a visual image as follows. A thermal equilibrium is assumed to be present wherein we have

$$W_{abs} = W_{cd} + W_{cv} + W_{rad}$$

where W_{cv} is the heat convected from the surface to the air, and W_{rad} is the heat lost by the surface due to radiation. The radiation flux is given by

$$W_{rad} = \epsilon_0 \sigma (T_s^4 - T_{amb}^4)$$

where T_s is the surface temperature, which can be obtained from the thermal image, and T_{amb} is the temperature of the air near the surface of the object, which is assumed to be known. Therefore W_{rad} can be obtained from the thermal image. The convective heat transfer is given by

$$W_{cv} = h(T_s - T_{amb})$$

where h is the average convection heat transfer coefficient. This coefficient can be estimated (see [114]) and we will assume it to be known. Thus W_{cv} can be obtained from the thermal image as well. The amount of heat absorbed by the surface depends on the amount of irradiance it receives. In [114] it was assumed that all of the irradiance was produced by the sun, and all of this irradiance was concentrated in the visible spectrum. Thus

$$W_{abs} = W_i \cos\theta_i \alpha_s$$

where θ_i is the angle between the surface normal vector and the illumination direction, and α_s is the solar absorptivity of the surface. W_i is the incident solar radiation (which can be determined, on a sunny day, from knowledge of the date and the time, and the solar insolation for that day. The values of $\cos\theta_i$ and α_i are assumed to be obtainable from the visual image (see details in [114]), so that W_{abs} can be determined from the visual image. We can now form the feature $\frac{W_{cv}}{W_{abs}}$. We can also compute the value of the conduction heat flux using

$$W_{cd} = W_{abs} - W_{cv} - W_{rad}$$

This heat flux can be used to estimate the thermal conductivity of the object, which may be useful in image segmentation or object recognition. Note that the values of these two derived features depend on both the thermal and visual images. Hence we have fused the two images in deriving the desired quantity. Note that this fusion is an algebraic one, and that there is no relative weighting of the thermal or visual data. Both are required in order to obtain the desired quantities.

4.6 EXAMPLES OF STRONGLY COUPLED FUSION IN THE VISION LITERATURE

There have been a number of examples of vision algorithms in the computational vision literature which utilize what can be described as strongly coupled data fusion. We review some of these here.

The most common application of strongly coupled methods use the Kalman filter. Kalman filter based data fusion algorithms are strongly coupled as they involve adaption of the prior model, through updating of the estimate of mean and covariance of the state variable. An example of the application of Kalman filter techniques to data fusion can be found in the work on depth-from-motion done by Matthies *et al* [105]. The operation of the Matthies *et al* feature based depth estimation procedure is as follows. A set of feature points $\{x\}$ are derived from an image (e.g. "edges"). These features are then tracked as the camera moves in a known manner, in a fashion similar to active vision. The state vector for this method is the image feature position x_t (at the current time t) and the estimate of the depth of the feature d_t. It is assumed that the camera motion is known exactly (and is lateral translation) and that the errors in the feature position measurements are normally distributed with variance σ_e^2. The initial state vector (x, d) and state covariance P is given by $x_1 = \tilde{x}_1, d_1 = (\tilde{x}_1 - \tilde{x}_0)/T_1$ and

$$P_1^+ = \sigma_e^2 \begin{pmatrix} 1 & -1/T_1 \\ -1/T_1 & 2/T_1^2 \end{pmatrix}$$

where $\tilde{x}_0$ and $\tilde{x}_1$ are the measured feature positions for the first two time steps, and T_1 is the camera translation. The state update equations are:

$$u_t^- = \begin{pmatrix} x_t^- \\ d_t^- \end{pmatrix} = \begin{pmatrix} 1 & -T_t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1}^+ \\ d_{t-1}^+ \end{pmatrix} = \Phi_t u_{t-1}^+$$

The covariance (prior model) update is

$$P_t^- = \Phi_t P_{t-1}^+ \Phi_t^T$$

The + superscript indicates that the variable includes the effect of the current measurement and the - superscript indicates that the variable does not include the effect of the current measurement. The information from the current measurement is embedded into the state and state covariance estimates as follows:

$$P_t^+ = ((P_t^-)^{-1} + S)^{-1}$$

$$u_t^+ = u_t^- + K(\tilde{x}_t - x_t^-)$$

where

$$S = \frac{1}{\sigma_e^2} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

and where the Kalman filter gain K is

$$K = \frac{1}{\sigma_e^2} P_t^+ \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Note that the covariance update does not actually depend on the measured data, but is deterministic.

If we assume that the measurements $\tilde{x}_t$ are independent with equal variance σ_e^2 one finds that the error in the depth has a decreasing variance over time given by:

$$\sigma_F^2(t) = \frac{12\sigma_e^2}{t(t+1)(t+2)}$$

This technique is strongly coupled because of the adaption of the prior model through the updating, dependent on the measured data, of the state variable (which is the mean of $p(f)$, the prior model). It is recurrent since the prior model depends on the previous output (the estimate u). The dynamics of this recurrence are made explicit in the Kalman filter method, however, and the convergence properties of such algorithms have been well studied.

A related approach was proposed by Szeliski [146]. This method used a Bayesian or energy function minimization approach to estimating motion from depth information. The technique assumes that the true depth values u are Gaussian random variables, with zero mean and covariance P (i.e. the surface has a random shape), and the depth measurements d are related to the actual values through the following linear transformation:

$$d = Hu + r$$

where H is a constant, sparse, matrix, and r is a Gaussian distributed random variable with zero mean and covariance R. Based on a sparse set of depth measurements at some time t_1, an estimate of the depth map u_1 can be determined with:

$$\hat{u}_1 = (P_0^-1 + H^T R^{-1} H)^{-1} H^T R^{-1} d$$

where R and H are assumed to be known *a priori* and P_0 is the current estimate (if it is not known *a priori*) of the covariance of the depth function. If the covariance is not known *a priori* it can be updated from P_0 as follows:

$$P_1 = (P_0^{-1} + H^T R^{-1} H)^{-1}$$

Note that this update does not depend on any measurements, and could be precomputed. Therefore the fusion algorithm that is to be described is not strongly coupled to the extent that the prior model is not affected by any measurements or external modules (even though it does change with time, it changes in a predetermined fashion). The strong coupling arises through recurrent adaption of the image formation model, as seen below.

Suppose the surface that we are viewing is moving. Then the image coordinates of the points on the surface will change. We can undo this effect by transforming the new depth values and image coordinates with a geometric transformation:

$$p_2 = T(p_2', \Theta)$$

where p_2' is the measured coordinates and p_2 are the transformed versions. T is the transformation and Θ is the set of parameters of this transformation (typically rotations and translations). Now, we do not know Θ as these are the quantities that we wish to determine by some computation on the depth values. Szeliski proposes the following procedure for determining Θ. Let d_2 be the transformed depth values, and let H_2, and R_2 be similarly transformations of H and R. We can find Θ as that which maximizes the following energy:

$$E(d_2(\Theta)) = \log|2\pi R_2| + (d_2 - H_2\hat{u}_1)^T R_2^{-1}(d_2 - H_2\hat{u}_1)$$

This effectively determines the likelihood that the new points d_2 comes from the (interpolated or estimated) surface $\hat{u}_1$. The minimization process is highly nonlinear and involves the joint minimization over six variables. One can see that the algorithm is strongly coupled as the determination of Θ depends on the specification of $\hat{u}$, which depends on previous estimates of Θ. That is, as the camera moves, we determine Θ given the previous estimate of the depth map. Once Θ is determined we can obtain an new estimate of the depth map from the properly transformed data d. We can use this new depth map to compute Θ for a new camera movement.

Szeliski proposes an extension to the above approach which uses the estimates of the depth at both times, $\hat{u}_1$ and $\hat{u}_2$. Here we are trying to find the value of Θ that makes it most likely that the data points p_1 and p_2 came from the same smooth surface. This extension uses the fact that the new depth measurements must come from a distribution, conditioned on the initial depth estimates $\hat{u}_1$, with zero mean and covariance $H_2 P_1 H_2^T + R_2$ (whereas in the above method the depth measurements were assumed to come from a distribution, unconditioned on $\hat{u}_1$, with zero mean and covariance R_2). Thus one needs to minimize the same energy function as above, save for replacing R_2 with $H_2 P_1 H_2^T + R_2$. As Szeliski points out, performing this minimization is difficult because of the large dimensionality of P_1 (same size as the image, and is not sparse). He gives the following approximation to the energy function which is more tractable computationally:

$$E(d_2(\Theta)) = (d_2 - H_2\hat{u}_1)^T R_2^{-1}(d_2 - H_2\hat{u}_2)$$

There is some coupling in this approach between the computation of Θ and the computation of $\hat{u}_2$, since $\hat{u}_2$ depends on R_2 and H_2.

The above two methods (except for the determination of the Θ parameters in Szeliski's algorithm) make the assumption that the relationship between the desired parameter and the measured data is linear. This is most often not the case, and the Kalman filter must be extended if it is to be used. Ayache and Faugeras [4] have proposed a method, based on the Extended Kalman Filter [50],

for fusing a number of noisy measurements to provide an estimate of the 3D position of objects in space. In this work Ayache and Faugeras use linear approximations to the function relating retinal coordinates to 3D world coordinates and Kalman filtering to attack some problems of weakly fusing visual information, particularly for stereo, assuming that the correspondence has been solved. In these models the uncertainties are due to lattice spacing errors in the two images; the probability distribution for the point in space is then approximated as Gaussian and the standard deviation is calculated. Kalman filtering techniques are then used get the optimal estimate for the position of the point in space.

The details of this approach are as follows. Consider the problem of determining the 3D location ($OM = (x, y, z)^T$) of a point M relative to the origin O of a reference frame from the coordinates $(u_1, v_1), (u_2, v_2)$ of the image of M in the image planes of two cameras. Assume that the correspondence problem has been solved, so that we can determine the (u_i, v_i)'s which correspond. For each image we get the following constraints on the position vector OM:

$$f(x, a) = \begin{pmatrix} (I_3 \cdot OM + l_{34})u - I_1 \cdot OM - l_{14} \\ (I_3 \cdot OM + l_{34})v - I_2 \cdot OM - l_{24} \end{pmatrix} = 0$$

where x represents the measurements $((u, v))$, a represents the parameters to be estimated (OM), and $I_1, I_2, I_3, l_{14}, l_{24}$, and l_{34} are known geometric quantities capturing the projection of points in space to the image plane.

The relationship between the desired parameter vector OM and the measurements u, v is a nonlinear one. In order to use a linear estimator for a we must linearize $f(x, a)$. This is done by expanding $f(x, a)$ in a Taylor's series about estimates, a^* and x, of a and x', where x' is the noise-free measurement ($x = x' + \epsilon$, where ϵ is zero mean Gaussian noise with covariance Λ) and retaining only terms of linear order. Thus we get

$$f(x', a) = 0 \approx f(x, a^*) + \frac{\partial f}{\partial x}(x, a^*)(x' - x) + \frac{\partial f}{\partial a}(x, a^*)(a - a^*)$$

This can be rearranged to yield the linear transformation

$$y = Ma + u$$

where y is a "measurement", $y = -f(x, a^*) + \frac{\partial f}{\partial a}(x, a^*)a^*$, u is a "noise" term, $u = \frac{\partial f}{\partial x}(x, a^*)\epsilon$ whose statistics are known ($E(u) = 0, E(uu^T) = W = \frac{\partial f}{\partial x}(x, a^*)\Lambda\frac{\partial f}{\partial x}(x, a^*)^T$), and M is an "observation" matrix, $M = \frac{\partial f}{\partial a}(x, a^*)$. Both y and M are known (since we know the form of $f(x, a)$). Thus we have reduced the problem to that of linear estimation of a. We can use the Kalman filter techniques described in chapter 2 to estimate a. This was the approach taken by Ayache and Faugeras in their algorithm. This algorithm has the nice property of being able to use information about the uncertainty of the measurements (u, v) to provide a measure of the uncertainty in the estimated parameter (the OM vector). Information from multiple camera positions is fused through the application of the Kalman filter. This algorithm was extended (in [4]) to estimate geometric parameters other than positions; the configuration of lines and planes in space can be obtained in a similar manner. The method was also extended to allow the estimation, or the improvement in an estimate, of the relative position and orientation of a camera as it moves about. This improvement comes about through a fusion of measurements about the configuration of points, lines, or planes in space, obtained, using the above techniques, from the camera images at the different locations. This fusion is performed using a Kalman filter process as well.

Geiger and Yuille [48] describe a method for strongly coupling shape from stereo with shape from controlled movements. For controlled movements, such as eye movements or head rotation, the matching problem (for each eye) can be solved by tracking, but large movements are required to get an accurate estimate of depth. Geiger and Yuille propose using crude depth estimates from small controlled movements to guide the stereo matching which then yields the correct depth. The system is strongly coupled since the information from the eye-movements directly affects the correspondence between the two eyes. An important aspect of this work is the precise modeling of the errors inherent in the controlled eye movement process.

Thus the system estimates the depth from eye movements but also gives an estimate of the reliability of this estimate. This system does not need *a priori* assumptions, such as the smooth surface assumption, and gives correct results when it is violated, for example by transparent surfaces. The system is also able to distinguish between the edges due to sharp boundaries, such as knife edges, and those due to smooth boundaries where the surface turns smoothly away from the viewer, as for a sphere. The key point is that for smooth boundaries the edges in the two eyes correspond to different points on the object, and the system is sensitive enough to detect this. More details on this algorithm are provided in chapter 6.

Another example of strong coupling is the work by House [67] on modeling the visual systems of toads and frogs. The visual systems of these animals have been extensively studied by Collett and coworkers [38]. Frogs have no conjunctive eye movements, and hence cannot perform binocular stereo, but can estimate depth in each eye separately using accommodation (depth from focus). House describes a cooperative algorithm that combines accommodation with stereo to influence the stereo matching and give a depth result. It is important to distinguish this from a weak method for combining stereo and accommodation. In such a method accommodation would not influence the stereo matching. The depth from accommodation and the depth from stereo would be computed separately and then combined using the estimates of their relative errors.

Singh [142] considers the fusion of motion information obtained from modules that use two different types of constraints, convection constraints and neighborhood constraints. A convection constraint is one in which there is some invariant property of the image as time progresses. For example, the motion analysis procedure of Horn and Schunk [65] assumes that the intensity of a moving patch does not change. A neighborhood constraint is one which is based on the spatial distribution of certain image parameters. This includes smoothness constraints, among others.

The form of the convection constraint assumed by Singh in his

motion analysis algorithm is as follows. A matching strength surface, $M(\delta x, \delta y, t)$ of extent $N \times N$, representing the likelihood of the inter-image displacement of a feature, is computed for each pixel in an image using the following formula

$$M(\delta x, \delta y, t) = \left(\sum_{i=-n}^{n} \sum_{j=-n}^{n} (I_0(x+i, y+j) - I_t(x+\delta x+i, y+\delta y+j))^2\right)^{-1}$$

A time average matching strength is determined by:

$$\hat{M}(\delta x, \delta y) = M(\delta x, \delta y, -1) + M(\delta x, \delta y, 0) + M(\delta x, \delta y, +1)$$

From this time average matching strength two possible displacement constraints are determined as the eigenvectors of the "Inertia" matrix S of the matching field $M(\delta x, \delta y)$. These eigenvectors can be interpreted as the axes about which the moment of inertia of the matching strength field is minimum and maximum. These eigenvectors provide two possibilities for a one-dimensional constraint on the flow vector. That is, we have two constraints on the displacement $(\delta x, \delta y)$ of the form

$$\mathrm{L} : a\delta x + b\delta y + c = 0$$

Let us call these two linear constraints L_1 and L_2. We can associate a confidence with each of these linear constraints by taking as the confidence the inverse of the (normalized) moment of inertia of the matching strength field about the constraint line.

Taken by themselves, the convection constraints could determine the flow vector by an algebraic, class II, weakly coupled fusion process that would correspond to finding the intersection of the two constraint lines. The confidences would not have any effect on the solution, since both constraints are absolutely needed to obtain a unique solution. If we had additional constraints we could use the confidences in a class III weakly coupled fusion algorithm. These extra constraints, in Singh's approach comes from the neighborhood constraints.

The neighborhood constraints in Singh's method are obtained as follows. The optical flow vectors (or image displacements) in a

window about each pixel in the image are mapped into the $\delta x, \delta y$ space, and a mass is associated with each point proportional to a Gaussian function of the distance between the point in the window and the center of the window. Thus points closest to the image pixel in question are weighted more heavily than are points further away. In this fashion we obtain a "mass" function, analogous to the matching strength function encountered in the determination of the convection constraints. We can compute the principle axes of this mass distribution using the same techniques as in the case of the convection constraints, and hence we can define two linear constraints on the image displacements. These have the same form as the convection constraints, and we will call them L_3 and L_4. We can associate confidences with these two constraints by taking as the confidence the inverse of the (normalized) moment of inertia of the mass distribution about the constraint line. Note that the neighborhood constraint depends on the quantity that we are trying to solve for, the displacement field, much in the same manner that smoothness constraints, or any *a priori* constraint on the solution, is a function over the space of possible solutions. In this case this "*a priori* " constraint is quite simple, being a linear constraint. This results in a simple, deterministic, algorithm for finding the solution which fuses the constraints.

The approach taken by Singh to fuse the four constraints is to find the displacements that minimize the following weighted (by the confidence values) sum of squares of the four constraints. If all of the constraints were satisfied (which is unlikely in general) this sum would be zero.

$$E = (1-\psi^2)((L_1C_1)^2 + (L_2C_2)^2) + \psi^2((L_3C_3)^2 + (L_4C_4)^2)$$

The quantity ψ is a factor which weights the relative importance of the neighborhood constraints versus the convection constraints. If ψ is small there is little dependence on the neighborhood constraints. Since the constraints are simple (they are linear) we can solve for the minimum exactly (if it exists). This is given by [142]

$$\delta x = \frac{t*r - q*s}{p*q - r^2}$$

$$\delta y = \frac{s*r - t*p}{p*q - r^2}$$

where

$$p = (1-\psi^2)(C_1^2a_1^2 + C_2^2a_2^2) + \psi^2(C_3^2a_3^2 + C_4^2a_4^2)$$

$$q = (1-\psi^2)(C_1^2b_1^2 + C_2^2b_2^2) + \psi^2(C_3^2b_3^2 + C_4^2b_4^2)$$

$$r = (1-\psi^2)(C_1^2a_1b_1 + C_2^2a_2b_2) + \psi^2(C_3^2a_3b_2 + C_4^2a_4b_2)$$

$$s = (1-\psi^2)(C_1^2a_1c_1 + C_2^2a_2c_2) + \psi^2(C_3^2a_3c_2 + C_4^2a_4c_2)$$

$$r = (1-\psi^2)(C_1^2c_1b_1 + C_2^2c_2b_2) + \psi^2(C_3^2c_3b_2 + C_4^2c_4b_2)$$

In Singh's algorithm the convection constraints are precomputed and remain constant, independent of the solution. Initially, however, we have no idea of what the displacement field is, and so cannot obtain a meaningful neighborhood constraint. Thus Singh uses an iterative method in which the factor ψ is initially taken to be 0, so that the solution depends only on the convection constraints, and then raises the value of ψ as a hopefully accurate displacement field begins to be derived. As with all recurrent strongly coupled approaches one must be concerned with the convergence of the algorithm. Singh claims that in practice the algorithm seems to converge to reasonable values but does not provide any analysis of the convergence. Presumably there will be, at the very least, pathological image flow fields for which the algorithm will either converge to an incorrect value, or not converge at all.

It is evident that Singh's fusional method is a strongly coupled one since it effectively alters the *a priori* probability of the solution (as measured by the neighborhood constraint) as the solution process proceeds. The convection constraint plays the role of the image formation model in our Bayesian fusional paradigm, as it expresses how the displacement field manifests itself in the image sequence (which is the data being fed into the algorithm). This algorithm is an interesting example as it describes the embedding of constraints, in essentially a Bayesian manner, that are not Gaussian or quadratic. It also is an example of the adaptive nature that is the signature of recurrent strongly coupled fusion techniques.

The work that is closest to the philosophy of this text is that of Poggio, Gamble and Little [123] and that of Chou and Brown [30]. These approaches are essentially the same as the coupled Markov Random Field methods described in chapter 3.

Poggio et al [123] use a more complex energy term involving the prior expectation of the line process field. The energy function in their approach is

$$U_i(f_i;l) = \sum_j (f_i - f_j)^2(1 - l_i^j) + \beta V_C(l_i^j)$$

where l_i^j is a binary line process element between lattice sites i and j. The term $V_C(l_i^j)$ embeds prior information about the likelihood of different configurations of the line process field. In their method strong coupling is obtained (apart from that induced by the coupling of the f and l fields) by letting V_C be dependent on some independent sensory module. They give an example of using a natural constraint, that changes in physical properties of object surfaces usually produce large gradients in image brightness, to alter V_C. This is done by letting

$$V_C(l_i^j, b_i^j) = b_i^j(1 - l_i^j)$$

where b_i^j is the gradient in the image brightness between lattice sites i and j. This method can be classified as a feedforward prior model adaption fusional algorithm. The prior constraint in the above method is embedded through the energy term incorporating V_C. Since this term depends on an measurement (the image gradient b) independent of the (f, l) module, there is strong coupling with adaption of the prior model.

The work of Chou and Brown [30] is also based on an application of coupled Markov Random Fields. Their goal was to integrate intensity information with a depth information to produce a segmentation of the depth map. The energy function they use is the following:

$$U(f,l|g,O) = \sum_{c\in C} V_c(f,l) + T\left(\sum_{s\in S} \frac{(f_s - g_s)^2}{2\sigma_s^2} - \sum_{d\in D} \log(\lambda_d(l_d))\right)$$

where f is the computed depth map, g is the measured depth, l is the computed depth discontinuity field, and O is the measured intensity. $V_C(f,l)$ is the energy of a given "clique" or configuration of discontinuities l, with the depth map f. A simple example of a suitable V_C is $V_C(f,l) = (1 - l_i^j)(f_i - f_j)^2$. C is the space of all possible "cliques" with respect to a neighborhood system Γ (e.g. in a one dimensional implementation Γ could be all pairs of adjacent lattice points), S is the set of lattice sites at which depth values are available, D is the set of lattice sites at which discontinuities can be placed (e.g. halfway between the elements of S). T is a temperature parameter and $\lambda_d(l_d)$ is the likelihood ratio of there being a discontinuity at site d given the intensity data O at site d.

This fusion method is clearly strongly coupled, as the intensity data and the computed discontinuities are used to compute the depth map and the discontinuity map (the segmentation). It is also clearly a recurrent approach due to the coupling between the depth map and the discontinuity field.

As for all recurrent fusional methods the dynamical aspects of the implementation are important for determining the convergence properties of the algorithm. Chou and Brown present a novel way of performing the minimization of the above energy function, a method they call Highest Confidence First (HCF). This approach is claimed to have computational advantages over iterative relaxation schemes. The reader is referred to [30] for more details on the HCF algorithm.

4.7 SUMMARY

- Applications of data fusion include reducing uncertainty, getting enough information to provide a unique solution, and modifying possibly invalid prior constraints.
- We classify fusional methods as being either weakly or strongly coupled, where the distinction is made by considering whether or not the operation of the modules are independent of each

other.

- Weakly coupled fusion is characterized by the combination of the outputs of independent information sources. Each module operates independently of the others (although the modules may themselves perform fusional operations).
- We divide the class of weakly coupled algorithms into a number of subclasses. The distinctions between these sub classes of weakly coupled fusion methods serve to point out that there are different reasons for performing sensory fusion and that different methods are called for in each case.
- The methods in class I are used when the goal is merely to reduce the uncertainty in a desired value, or when one wishes to reduce reliance on the assumptions of a given sensory module.
- The methods in class II are indicated when the modules available do not provide unique solutions.
- The methods of class III are typically of use when one wishes to combine the outputs of sensory modules that, by themselves, do not provide unique or stable outputs, and at the same time weight the information from the component modules according to their relative reliabilities in order to minimize the uncertainty in the fused output.
- Strongly coupled data fusion involves alteration of the prior constraints (either the image formation model, prior model, or system model) used by a information processing module by the output of another module or set of modules.
- Recurrent strong coupling occurs when the output of a given module is fed back, through some path, to affect its own prior constraints. Common examples of recurrent strong coupling are to be found in some Kalman filter based fusional methods, and in coupled Markov Random Field based methods.

Chapter 5

Data Fusion Applied to Feature Based Stereo Algorithms

5.1 INTRODUCTION

In this chapter we describe a theoretical formulation for stereo (this was first proposed by Yuille, Geiger and Bülthoff in [169]) in terms of the Bayesian approach to vision outlined in chapters 2 and 3, in particular in terms of coupled Markov Random Fields. We show that this formalism is rich enough to contain most of the elements used in standard stereo theories.

This formulation enables us to integrate the depth information obtained using different types of matching primitives, or from different vision modules.

The fundamental issues of the binocular stereo process are: (i) what primitives are matched between the two images, (ii) what *a priori* assumptions are made about the scene to determine the matching and thereby compute the depth, and (iii) how is the geometry and calibration of the stereo system determined. In this text we assume that (iii) is solved, and so the corresponding epipolar lines between the two images are known. Thus we use the epipolar line constraint for matching. Some support for this assumption is given by the work of Bülthoff and Fahle [24], described in more detail in [169].

Our framework permits combining cues from different matching primitives to obtain an overall perception of depth. These primitives can be weighted according to their robustness. For example, depth estimates obtained by matching intensity are sometimes unreliable since small fluctuations in intensity (due to illumination or detector noise) can lead to large fluctuations in depth, and hence are less reliable than estimates obtained from the matching of edge features. The formalism can also be extended to incorporate information from other depth modules, such as the depth provided by a shape from shading module. Such a strongly coupled approach is described in the next chapter. The energy function used as the basis of this framework was initially described in [170], but without the statistical tools needed to analyze it.

Unlike previous theories of stereo which first solved the correspondence problem and then constructed a surface by interpolation (e.g. Grimson's approach [57]), our theory proposes combining the two stages. The correspondence problem is solved to give the disparity field which best satisfies the *a priori* constraints. Our model involves the interaction of several processes and is fairly complex. We will introduce it in three stages at different levels of complexity.

At the first level features (such as edges) are matched, using a binary matching field V_{ia} determining which features correspond. In addition smoothness is imposed on the disparity field $d(\vec{x})$ which is related to the depth of the surface from the fixation plane. In this case the correspondence problem, i.e. determining the V_{ia}, is solved to give the smoothest possible disparity field. It is related to the work by Yuille and Grzywacz [165] on motion measurement and correspondence, and, in particular, to work on long-range motion. We show later in the chapter that cooperative stereo algorithms [37, 99] are closely related to this theory.

At the second level of complexity we add line process fields $l(\vec{x})$ (which represents depth discontinuities) [51] to break the surfaces where the disparity gradient becomes too high. This level is related to theories based on the disparity gradient limit [121, 129].

The third level introduces additional terms corresponding to matching image intensities. Such terms are used in the theories of Gennert [52] and Barnard [8] which, however, do not have line process fields or matching fields. A psychophysical justification for intensity matching is given by the work of Bülthoff and Mallot [25]. Thus our full theory is expressed in terms of energy functions relating the disparity field $d(\vec{x})$, the matching field V_{ia}, and the line process field $l(\vec{x})$.

As described in chapter 3, with the use of standard techniques from statistical physics, we can eliminate certain fields and obtain effective energies for the remaining fields [46, 48]. As discussed in [166] (following Lumsdaine *et al* [93]) this can be interpreted as computing marginal probability distributions. We use this to show that several existing stereo theories are closely related to versions of our model.

These mean field techniques also suggest novel algorithms for stereo computation. We argue that these algorithms incorporate constraints about the set of possible matches better than previous algorithms. They can also be directly related [166] to analog methods for solving the traveling salesman problem. Moreover the greater empirical success of the elastic net algorithm [42] compared with the Hopfield and Tank method [61] strongly suggests that our novel stereo algorithms will be more successful than some existing algorithms.

This model can be related [169] to some psychophysical experiments [26, 24] in which perceived depth for different matching primitives and disparity gradients are precisely measured. Their results suggest that several types of primitive are used for correspondence, but that some primitives are better than others. Our model is in good general agreement with the data from these experiments.

5.2 THE BAYESIAN APPROACH TO STEREO VISION

5.2.1 THE MATCHING PROBLEM

The input to any binocular stereo system is a pair of images. The task is to match primitives of the two images, thereby solving the correspondence problem. The depth of objects in the scene can then be determined by triangulation, assuming the orientations of the cameras (and other camera parameters) are known. In many stereo theories the disparity, the relative distance between matched features, is first computed. The depth of the feature from the fixation point is then, to good approximation, linearly dependent on its disparity.

There are several choices of matching primitives. Some theories use features such as edges or peaks in the image intensity (e.g., Marr and Poggio [99]; Pollard, Mayhew and Frisby [121]; Prazdny [129] while others match the image intensity directly (e.g., Barnard [8]; Gennert [52]). Yet another class of theory acts on the Fourier components of the images (e.g., Sanger [136]; Jepson and Jenkin [74]) and hence is particularly sensitive to texture. It is unclear which primitives the human visual system uses. Current psychophysical research [26, 24] suggests that at least edges and image intensity are used as primitives.

It is desirable to build a stereo theory that is capable of using all these different types of primitives. This will allow to reduce the

complexity of the correspondence problem and will enhance the robustness of the theory and its applicability to natural images. But not all primitives are equally reliable, however. A small fluctuation in the image intensity might lead to a large change in the measured disparity for a system which matches intensity. Thus image intensity tends to be less reliable than features such as edges.

Some assumptions about the scene being viewed are usually necessary to solve the correspondence problem. These can be thought of as natural constraints and, as we have seen in previous chapters, are needed because of the ill-posed nature of vision. There are two types of assumption: (i) assumptions about the matching primitives, i.e., that similar features match (*compatibility constraint*), and (ii) assumptions about the surface being viewed (*continuity constraint*). For (ii) one typically assumes that either the surface is close to the fixation point (disparity is small) or that the surface's orientation is smoothly varying (disparity gradient is small) with possible discontinuities.

Our theory requires both assumptions but their relative importance depends on the scene. If the features in the scene are sufficiently different then assumption (i) is often sufficient to obtain a good match. If all features are very similar, assumption (ii) is necessary. We require that the matching is chosen to obtain the smoothest possible surface, so interpolation and matching are performed simultaneously (the next section formalizes these ideas).

5.2.2 THE FIRST LEVEL: MATCHING FIELD AND DISPARITY FIELD

The basic idea in the simplest version of the binocular stereo algorithm is that there are a number of possible primitives that could be used for matching and that these all contribute to a disparity field $d(x)$. This disparity field exists even where there is no source of data. The primitives we will consider here are image features, such

as edges in image brightness. Edges typically correspond to object boundaries, and other significant events in the image. Other primitives, such as peaks in the image brightness or texture features, can also be added. We will describe the theory for the one-dimensional case.

We assume that the edges and other features have already been extracted from the image in a preprocessing stage. The matching elements in the left eye consist of the features x_{i_L}, for $i_L = 1, ..., N_l$. The right eye contains features x_{a_R}, for $a_R = 1, ..., N_r$. We define a set of binary matching elements $V_{i_L a_R}$, the matching field, such that $V_{i_L a_R} = 1$ if point i_L in the left eye corresponds to point a_R in the right eye, and $V_{i_L a_R} = 0$ otherwise. A *compatibility field* $A_{i_L a_R}$ is defined over the range $[0,1]$. For example, it is 1 if i_L and a_R are compatible (i.e. features of the same type), 0 if they are incompatible (an edge cannot match a peak).

We now define a cost function $E(d(x), V_{i_L a_R})$ of the disparity field and the matching elements. We will interpret this in terms of Bayesian probability theory in the next section. This will suggest several methods to estimate the fields $d(x), V_{i_L a_R}$ given the data. A standard estimation procedure is to minimize $E(d(x), V_{i_L a_R})$ with respect to $d(x), V_{i_L a_R}$.

$$E(d(x), V_{i_L a_R}) = \sum_{i_L, a_R} A_{i_L a_R} V_{i_L a_R} (d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + \lambda \{ \sum_{i_L} (\sum_{a_R} V_{i_L a_R} - 1)^2 + \sum_{a_R} (\sum_{i_L} V_{i_L a_R} - 1)^2 \} + \gamma \int_M (Sd)^2 dx \quad (5.1)$$

The first term gives a contribution to the disparity obtained from matching i_L to a_R. The third term imposes a smoothness constraint on the disparity field imposed by a smoothness operator S.

The second term encourages features to have a single match. This can be avoided by requiring that each column and row of the matrix

$V_{i_L a_R}$ contains only one 1. An extension to this approach by Yuille *et al* [171] allows the unmatching, at a cost, of matched features. In Section 5.3 we will argue that it is better to impose constraints in this way, hence the second term will not be used in our final theory. However we will keep it in our energy function for the present since it will help us relate our approach to alternative theories.

Minimizing the energy function with respect to $d(\vec{x})$ and $V_{i_L a_R}$ will cause the matching which results in the smoothest disparity field.

The coefficient γ determines the amount of *a priori* knowledge required. If all the features in the left eye have only one compatible feature in the right eye then little *a priori* knowledge is needed and γ may be small. If all the features are compatible then there exists a matching ambiguity for which the *a priori* knowledge is needed to resolve, requiring a larger value of γ and hence more smoothing. In Section 5.5 we show that this gives a possible explanation for some psychophysical experiments.

The theory can be extended to the two dimensional case in a straightforward way. The matching elements $V_{i_L a_R}$ must be constrained to only allow for matches that use the epipolar line constraint. The disparity field will have a smoothness constraint perpendicular to the epipolar line which will enforce figural continuity. Note that the epipolar constraint is a physical constraint (not a natural constraint) as it depends on the laws of geometry and propagation of light. It should be noted that, although the form of the epipolar constraint is usually valid (since the laws of mathematics and physics associated with the epipolar constraint are presumably usually valid), some of its parameters may be imperfectly known (such as the relative angle between the optical axes of the cameras) and hence may produce erroneous results. The smoothness constraint is a natural constraint (and not a physical constraint) as it is not based on laws of physics or mathematics, but upon the subjective opinion that surfaces (and hence disparity fields) are smooth.

We must choose a form for the smoothness operator S. Marr

[97] proposed that, to make stereo correspondence unambiguous, the human visual system assumes that the world consists of smooth surfaces. This suggests that we should choose a smoothness operator which encourages the disparity to vary smoothly spatially. In practice the assumptions used in Marr's two theories of stereo are somewhat stronger. Marr and Poggio I (the cooperative stereo algorithm [99]) encourages matches with constant disparity, thereby enforcing a bias to the fronto-parallel plane. Marr and Poggio II (the multiresolution stereo algorithm [99]) uses a coarse to fine strategy to match nearby points, hence encouraging matches with minimal disparity and thereby giving a bias towards the fixation plane.

An alternative approach is to introduce discontinuity fields which break the smoothness constraint, as is done in the next section. For these theories the experiments described in Section 5.5 are consistent with S being a first order derivative operator. This is also roughly consistent with Marr and Poggio's cooperative stereo algorithm [99]. We will therefore use $S = \partial/\partial x$ as a default choice for our theory.

5.2.3 THE SECOND LEVEL: ADDING DISCONTINUITY FIELDS

The first level theory is easy to analyze but makes the *a priori* assumption that the disparity field is smooth everywhere, which is false at object boundaries. There are several standard ways to allow smoothness constraints to break [14, 51, 113]. We introduce a discontinuity or line process field $l(x)$ represented by a set of curves C.

Introducing the discontinuity fields C gives an energy function

$$E(d(x), V_{i_L a_R}, C) = \sum_{i_L, a_R} A_{i_L a_R} V_{i_L a_R} (d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + \lambda \{ \sum_{i_L} (\sum_{a_R} V_{i_L a_R} - 1)^2 + \sum_{a_R} (\sum_{i_L} V_{i_L a_R} - 1)^2 \} + \gamma \int_{M-C} (Sd)^2 dx + M(C) \quad (5.2)$$

where smoothness is not enforced across the curves C. $M(C)$ is the cost for enforcing breaks, and is proportional to the length of the curves C. Once again the second term on the right hand side of (5.2) will not appear in the final version of the theory. As was mentioned in chapter 4, we can think of this approach as strongly fusing a stereo module with a discontinuity detection module. The feature matching influences the discontinuity detection and vice versa.

In the next section we will apply some of the mean field based methods described in chapter 3 to the computation of some of the properties of the above energy function, such as its minimum.

5.2.4 THE THIRD LEVEL: ADDING INTENSITY TERMS

The final version of the theory couples intensity based and feature based stereo. Psychophysical results by Bülthoff and Mallot (see Section 5.5) suggest that this is necessary. Our energy function becomes

$$E(d(x), V_{i_L a_R}, C) = \sum_{i_L, a_R} A_{i_L a_R} V_{i_L a_R} (d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + \mu \int \{L(x) - R(x + d(x))\}^2 dx + \lambda \{ \sum_{i_L} (\sum_{a_R} V_{i_L a_R} - 1)^2 + \sum_{a_R} (\sum_{i_L} V_{i_L a_R} - 1)^2 \} + \gamma \int_{M-C} (Sd)^2 dx + M(C)$$

If certain terms are set to zero in the above equation it reduces to previous theories of stereo. If the second and fourth terms are kept, without allowing discontinuities, it is similar to work by Gennert [52] and Barnard [8]. If we add the fifth term, and allow discontinuities, we get connections to a theory described in [167]. The third term will again be removed in the final version of the theory.

Thus the cost function reduces to well-known stereo theories in certain limits. It also shows how it is possible to combine feature and brightness data in a natural manner. In addition it can be modified to include monocular cues as is done in the next chapter.

A similar theory for integrating different cues for motion perception was proposed in [165], although it did not involve discontinuity fields.

5.2.5 THE BAYESIAN FORMULATION OF THE STEREO ALGORITHM

Given an energy function model one can define a corresponding statistical theory. If the energy $E(d, V, C)$ depends on three fields: d (the disparity field), V the matching field and C (the discontinuities), then (using the Gibb's distribution – see [119]) the probability of a particular state of the system is defined by

$$P(d, V, C|g) = \frac{e^{-\beta E(d,V,C)}}{Z}$$

where g is the data, β is the inverse of the temperature parameter and Z is the partition function (a normalization constant).

Using the Gibb's Distribution we can interpret the results in

terms of Bayes' formula

$$P(d, V, C|g) = \frac{P(g|d, V, C)P(d, V, C)}{P(g)}$$

where $P(g|d, V, C)$ is the probability of the data g given a scene expressed in terms of a disparity field d, a matching of features V, and a set of disparity discontinuities C. $P(d, V, C)$ is the prior model, or the *a priori* probability of the scene and $P(g)$ is the *a priori* probability of the data. Note that $P(g)$ appears in the above formula as a normalization constant, so its value can be determined if $P(g|d, V, C)$ and $P(d, V, C)$ are assumed known.

This implies that every state of the system has a finite probability of occurring. The more likely ones are those with low energy. This statistical approach is attractive because the β parameter gives us a measure of the uncertainty of the model through the temperature parameter $T = \frac{1}{\beta}$. At zero temperature ($\beta \to \infty$) there is no uncertainty. In this case the only state of the system that has nonzero probability, and hence probability 1, is the state that globally minimizes $E(d, V, C)$. In some nongeneric situations, however, there could be more then one global minimum of $E(d, V, C)$.

Minimizing the energy function will correspond to finding the most probable state, independent of the value of β. The mean field solution,

$$\bar{d} = \sum_{d,V,C} dP(d, V, C|g)$$

is more general and reduces to the most probable solution as $T \to 0$. It corresponds to defining the solution to be the mean fields, the averages of the f and l fields over the probability distribution. This enables us to obtain different solutions depending on the temperature parameter.

In this chapter we concentrate on using the mean quantities of the field (these can be related to the minimum of the energy function in the zero temperature limit). A justification for using the mean field as a measure of the fields resides in the fact that it represents the

minimum variance Bayes estimator (Gelb 1974). More precisely, the variance of the field d is given by

$$Var(d:\bar{d}) = \sum_{d,V,C} (d-\bar{d})^2 P(d,V,C|g)$$

where $\bar{d}$ is the center of the variance and the $\sum_{d,V,C}$ represents the sum over all the possible configurations of d, V, C. Minimizing $Var(d:\bar{d})$ with respect to all possible values of $\bar{d}$ we obtain

$$\frac{\partial}{\partial \bar{d}} Var(d:\bar{d}) = 0 \rightarrow \bar{d} = \sum_{d,V,C} dP(d,V,C)$$

This implies that the minimum variance estimator is given by the mean field value.

5.3 STATISTICAL MECHANICS AND MEAN FIELD THEORY

In this section we describe the application of the mean field based methods given in chapter 3 to the task of calculating the quantities we are interested in from the energy function. These will lead to novel stereo vision algorithms.

For the first level stereo theory, as described in section 5.2.2, it is possible to eliminate the disparity field to obtain an effective energy $E_{eff}(V_{ij})$ depending only on the binary matching field V_{ia}. The resulting algorithm is related to those obtained from cooperative stereo theories (e.g. [37, 99]). Alternatively, we can eliminate the matching fields to obtain an effective energy $E_{eff}(d)$ depending only on the disparity. We believe that the second approach is better since it incorporates the constraints on the set of possible matches implicitly rather than imposing them explicitly in the energy function (as the first method does).

Moreover it can be shown [166] that there is a direct correspondence between these two theories (with $E_{eff}(V_{ia})$ and $E_{eff}(d)$) and

the analog models for solving the traveling salesman problem proposed by Hopfield and Tank [61] and Durbin and Willshaw [42]. The greater empirical success of the Durbin and Willshaw algorithm suggests that the first level stereo theory based on $E_{eff}(d)$ will be more effective than the cooperative stereo algorithms.

We can also average out the line process fields or the matching fields or both for the second and third level theories. This leaves us again with a theory depending only on the disparity field.

Another approach is to use mean field theory methods to obtain deterministic algorithms for minimizing the first level theory $E_{eff}(V_{ia})$. These differ from the standard cooperative stereo algorithms and should be more be more effective (though not as effective as using $E_{eff}(d)$) since they can be interpreted as performing the cooperative algorithm at finite temperature thereby smoothing local minima in the energy function.

Our proposed stereo algorithms, therefore, consist of eliminating the matching field and the line process field by these statistical techniques leaving an effective energy depending only on the disparity field. This formulation will depend on a parameter β (which can be interpreted as the inverse of the temperature of the system). We then intend to minimize the effective energy by steepest descent while lowering the temperature (increasing β). This can be thought of as a deterministic form of simulated annealing [79] and has been used by many algorithms, for example [61, 42, 46]. It is also related to continuation methods [156].

5.3.1 AVERAGING OUT FIELDS

In the next few sections we provide the details on how one can, for the first and second level theories, average out fields to obtain equivalent, though apparently different, formulations. As discussed in [166] (following [93]) this can be interpreted as computing marginal

probability distributions.

AVERAGING OUT THE DISPARITY FIELD FOR THE FIRST LEVEL THEORY

We now show that, if we consider the first level theory, we can eliminate the disparity field and obtain an energy function depending on the matching elements V only. In the next section we will relate this to cooperative stereo algorithms.

The disparity field is eliminated by minimizing and solving for it as a function of the V [165]. Since the disparity field occurs quadratically this is equivalent to doing mean field over the disparity [119].

For the first level theory, assuming all features are compatible, our energy function becomes

$$E(d(x), V_{i_L a_R}) = \sum_{i_L, a_R} V_{i_L a_R}(d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + \mu \sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \mu \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2 + \lambda \int_M (Sd)^2 dx$$

with Euler-Lagrange equations

$$\lambda S^2 d(x) = \sum_{i_L, a_R} V_{i_L a_R}(d(x_{i_L}) - (x_{a_R} - x_{i_L}))\delta(x - x_{i_L})$$

The solutions of this equation are given by

$$d(x) = \sum_{i_L} \alpha_{i_L} G(x, x_{i_L})$$

where the $G(x, x_{i_L})$ are the Green function of the operator S^2, and

the α_{i_L} obey

$$\sum_{i_L} \alpha_{i_L}(\lambda \delta_{i_L a_R} + G(x_{i_L}, x_{a_R})) = \sum_{i_L} V_{i_L a_R}(x_{i_L} - x_{a_R})$$

Substituting this back into the energy function (assuming that each feature is matched precisely once, i.e. a *uniqueness constraint*) eliminates the disparity field and yields

$$E(V_{i_L a_R}) = \lambda \sum_{i_L, j_L}(\sum_{a_R} V_{i_L a_R}(x_{i_L} - x_{a_R}))(\lambda \delta_{i_L j_L} + G(x_{i_L}, x_{j_L}))^{-1} \times (\sum_{b} V_{j_L b_R}(x_{j_L} - x_{b_R})) + \lambda \sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \lambda \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2$$

This calculation shows that the disparity field is strictly speaking unnecessary as it does not appear in the above energy functional. The connection of this approach to cooperative stereo algorithms is discussed in the next section. A similar calculation [165] can be performed to show that Ullman's minimal mapping theory [150] was a special case of the motion coherence theory of Yuille and Grzywacz [165].

A weakness of the formulation of the stereo vision theory expressed in the minimization of the above energy, and the cooperative stereo algorithms related to it, is that the uniqueness constraints are imposed as penalties in the energy function, by the second and third terms on the right hand side. As mentioned earlier we believe it is preferable to use mean field theory techniques which enforce the constraints strictly.

AVERAGING OUT THE MATCHING FIELDS FOR THE FIRST LEVEL THEORY

We prefer an alternative way of writing the first level theory. This can be found by using techniques from statistical physics to average

out the matching field, leaving a theory which depends only on the disparity field.

The partition function for the first level system, again assuming compatibility between all features, is defined to be

$$Z = \sum_{V_{i_L a_R}, d(x)} e^{-\beta E(V_{i_L a_R}, d(x))}$$

where the sum is taken over all possible states of the system determined by the fields V and d.

It is possible to explicitly perform the sum over the matching field V yielding an effective energy for the system depending only on the disparity field d. Equivalently we could obtain the marginal probability distribution $p(d|g)$ from $p(d, V|g)$ by integrating out the V field [93].

To compute the partition function we must first decide what class of $V_{i_L a_R}$ we wish to sum over. We could sum over all possible $V_{i_L a_R}$ and rely on the $\lambda \sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \lambda \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2$ term to bias against multiple matches. Alternatively we could impose the constraint that each point has a unique match by only summing over $V_{i_L a_R}$ which contain a single 1 in each row and each column. We could further restrict the class of possible matches by requiring that they satisfied the ordering constraint[1].

For this section we will initially restrict that each feature in the left image has a unique match in the right image, but not vice versa. We could also allow no match, at some cost [171]. This simplifies the computation of the partition function, but we will relax it at the end of the section. The requirement of smoothness on the disparity field should ensure that unique matches occur. This is suggested by mathematical analysis of a similar algorithm used for an elastic network approach to the Traveling Salesman Problem [42].

[1] The ordering constraint requires that the spatial ordering along the epipolar line of matched features be the same in the left and right images

Since we are attempting to impose the unique matching constraint by restricting the class of V's the $\lambda \sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \lambda \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2$ terms do not need to be included in the energy function. We can now write the partition function as

$$Z = \sum_{V,d} \prod_{i_L} e^{-\beta\{\sum_{a_R} V_{i_L a_R}\left(d(x_{i_L}) - (x_{a_R} - x_{i_L})\right)^2 + \int_M (Sd)^2 dx\}}$$

For fixed i_L we sum over all possible $V_{i_L a_R}$, such that $V_{i_L a_R} = 1$ for only one a_R (this ensures that points in the left image have a unique match to points in the right image). This gives

$$Z = \sum_d \prod_{i_L} \{\sum_{a_R} e^{-\beta\left(d(x_{i_L}) - (x_{a_R} - x_{i_L})\right)^2}\} e^{-\beta \int_M (Sd)^2 dx}$$

This can be written using an effective energy $E_{eff}(d)$ as

$$Z = \sum_d e^{-\beta E_{eff}(d)}$$

where

$$E_{eff}(d) = -\frac{1}{\beta} \sum_{i_L} \log\{\sum_{a_R} e^{-\beta\left(d(x_{i_L}) - (x_{a_R} - x_{i_L})\right)^2}\} + \int_M (Sd)^2 dx$$

It is probably preferable, however, to impose symmetry between the two eyes. We do this by summing over states where the points in the right image have a unique match to points in the left image. This modifies the effective energy by the addition of a term $E_{aux}(d)$, where

$$E_{aux}(d) = -\frac{1}{\beta} \sum_{a_R} \log\{\sum_{i_L} e^{-\beta\left(d(x_{a_R}) - (x_{a_R} - x_{i_L})\right)^2}\}$$

Preliminary experiments [171] suggest that the symmetric energy function $E_{eff}(d) + E_{aux}(d)$ has fewer local minima that $E_{eff}(d)$ and is easier to compute.

Thus our first level theory of stereo can be formulated in this way without explicitly using a matching field. We are not aware, however, of any existing stereo theory of this form. Since it has formulated the matching constraints in computing the partition function we believe it is preferable to standard cooperative stereo algorithms.

AVERAGING OUT THE DISCONTINUITY FIELD FOR THE SECOND LEVEL THEORY

The second level theory includes a discontinuity field in addition to the matching field and the disparity field. The discontinuity and the matching fields are both binary and can be averaged out. In this section we will average out the discontinuity field, in the next section we will average out both fields. After averaging out the discontinuity field we will obtain a theory, depending only on the disparity field and the matching field, which is reminiscent of disparity gradient limit theories [121, 129].

We discretize the energy in (5.2) by replacing the discontinuity field C with a binary variable l_k and assuming a first derivative for the smoothness operator [51].

$$E[d,V,l] = \sum_{i_L,a_R} V_{i_L a_R}\{d(x_{i_L}) - (x_{a_R} - x_{i_L})\}^2$$
$$+\lambda[\sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2 + \sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2]$$
$$+\alpha\sum_k [d_k - d_{k-1}]^2(1-l_k) + \gamma l_k$$

where the l_k are binary elements which, when switched on $l_k = 1$, cut the smoothing constraints between the disparities d_k and d_{k+1}. The λ term enforces one and only one match. The index k scans all the pixels in the left image, while the indices i_L , a_R are only for feature pixels in the left and right images respectively. We can rewrite the λ term, up to an additive constant, as

$$\lambda[\sum_{i_L,a_R,b_R} V_{i_L a_R} V_{i_L b_R} + \sum_{i_L j_L a_R} V_{i_L a_R} V_{j_L a_R} - 2\sum_{i_L a_R} V_{i_L a_R}]$$

We can now integrate out the contributions from the l_k's to the partition function

$$Z = \sum_{d,V,l} e^{-\beta E[d,V,l]}$$

This can be written

$$Z = \prod_k \prod_{i_L} \prod_{a_R} \sum_d \sum_{V_{i_L a_R}} \sum_{l_k} exp\{-\beta[\alpha(d_k - d_{k-1})^2(1-l_k) + \gamma l_k]\}$$
$$exp\Big\{-\beta\{V_{i_L a_R}[d(x_{i_L}) - (x_{a_R} - x_{i_L})]^2$$
$$+\lambda[\sum_{b_R} V_{i_L a_R} V_{i_L b_R} + \sum_{j_L} V_{i_L a_R} V_{j_L a_R} - 4V_{i_L a_R}]\}\Big\}$$

where the product over k is over all the pixels and $V_{i_L a_R} = 0$ for any pixel without a feature. Performing the sum over the l_k

$$Z = \prod_i \prod_{a_R} \prod_{i_L} \sum_d \sum_{V_{i_L a_R}} (1 + e^{-\beta[\alpha(d_k - d_{k-1})^2 - \gamma]}) e^{-\beta\Gamma}$$

where Γ is

$$\Gamma = \Big\{V_{i_L a_R}[d(x_{i_L}) - (x_{a_R} - x_{i_L})]^2$$
$$+\lambda[\sum_{b_R} V_{i_L a_R} V_{i_L b_R} + \sum_{j_L} V_{i_L a_R} V_{j_L a_R} - 4V_{i_L a_R}] + \gamma\Big\}$$

This is of form

$$Z = \sum_d \sum_{V_{i_L a_R}} e^{-\beta E_{eff}(d,V)}$$

where the effective energy is

$$E_{eff}(d,V) = \sum_k \gamma - \frac{1}{\beta}\log(1 + e^{-\beta[\alpha(d_k - d_{k-1})^2 - \gamma]})$$
$$+ \sum_{a_R, i_L} V_{i_L a_R}\Big\{[d(x_{i_L}) - (x_{a_R} - x_{i_L})]^2$$
$$+\lambda[\sum_{b_R} V_{i_L a_R} V_{i_L b_R} + \sum_{j_L} V_{i_L a_R} V_{j_L a_R} - 4V_{i_L a_R}]\Big\}$$

We can now treat $E_{eff}(d,V)$ as a formulation of the second level theory with no discontinuity fields. Its relation to disparity gradient limit theories will be discussed in Section 5.4.

Although the discontinuity field has been eliminated from the effective energy it is still possible to obtain estimates of its value after $E_{eff}(V,d)$ has been minimized with respect to V and d. The necessary analysis has already been performed for the image segmentation case (Geiger and Girosi 1989) and yields

$$\bar{l}_k = \frac{1}{(1+e^{-\beta[\alpha(\bar{d}_k-\bar{d}_{k-1})^2-\gamma]})}$$

where $\bar{d}_k$ is the solution of the disparity field. Note that as $\beta \to \infty$ the discontinuity field l_k will tend to 0 or 1.

AVERAGING OUT THE MATCHING AND DISCONTINUITY FIELDS FOR THE SECOND LEVEL THEORY

It will usually be preferable to average out the discontinuity field and the matching fields simultaneously. This is possible because of the lack of interaction, or coupling, between them in the second level theory, see (2). It will yield an effective energy depending only on the disparity field.

We essentially combine the calculations of the previous two sections. This gives an effective energy

$$\begin{aligned} E_{eff}(d) = &-\frac{1}{\beta}\sum_{i_L}\log\{\sum_{a_R} e^{-\beta\left(d(x_{i_L})-(x_{a_R}-x_{i_L})\right)^2}\} \\ &-\frac{1}{\beta}\sum_{a_R}\log\{\sum_{i_L} e^{-\beta\left(d(x_{a_R})-(x_{a_R}-x_{i_L})\right)^2}\} \\ &-\frac{1}{\beta}\log(1+e^{-\beta[\alpha(d_k-d_{k-1})^2-\gamma]}) \end{aligned}$$

We can minimize with respect to d using steepest descent. The parameter β can be varied to allow for a deterministic annealing approach. Once again we can calculate the matching and discontinuity fields in terms of the disparity field.

AVERAGING OUT THE MATCHING AND DISCONTINUITY FIELDS FOR THE THIRD LEVEL THEORY

The third level theory differs from the second level theory only by the addition of intensity terms which are not strongly coupled to the matching and discontinuity fields. Thus to obtain the effective energy for this theory we merely need to add the intensity terms to the effective energy found in the previous section. This yields

$$\begin{aligned} E_{eff}(d) = &-\frac{1}{\beta}\sum_{i_L}\log\{\sum_{a_R} e^{-\beta\left(d(x_{i_L})-(x_{a_R}-x_{i_L})\right)^2}\} \\ &-\frac{1}{\beta}\sum_{a_R}\log\{\sum_{i_L} e^{-\beta\left(d(x_{a_R})-(x_{a_R}-x_{i_L})\right)^2}\} \\ &-\frac{1}{\beta}\log(1+e^{-\beta[\alpha(d_k-d_{k-1})^2-\gamma]}) \\ &+\mu\int\{L(x)-R(x+d(x))\}^2dx \end{aligned}$$

Again a deterministic annealing approach should yield good solutions to this problem.

5.3.2 DETERMINISTIC SOLUTIONS OF THE MEAN FIELD EQUATIONS

The mean field theory approach can also yield deterministic algorithms for theories including the binary matching elements (although we believe these algorithms will be inferior to methods which eliminate the matching fields for the reasons discussed in Section 5.1.2). For the related problem of image segmentation deterministic algorithms of this type seem to give equivalent results to stochastic algorithms, such as simulated annealing, and to run much faster [14].

Ideally we would like to compute the partition function of the system explicitly and then directly differentiate to obtain the mean

fields. However, computing the partition function is impossible in general. Instead we use a technique from statistical physics, *the mean field theory approximation*, which will give us a set of equations, *the mean field equations*, which must be satisfied by the mean fields. We can then use a deterministic algorithm, which can be thought of as a deterministic form of the Monte Carlo algorithm, to obtain a solution to the mean field equations. They may be several possible solutions to these equations, however, and we cannot guarantee to find the correct one.

A DETERMINISTIC MEAN FIELD ALGORITHM FOR THE LEVEL 1 THEORY

We first illustrate the mean field approximation technique for the energy function obtained from the first level theory by eliminating the disparity fields.

The energy function is of the form

$$E[V_{ia}] = \sum_{ijab} T_{ijab} V_{ia} V_{jb} + \sum_{ia} W_{ia} V_{ia}$$

where, for simplicity, we have replaced the indices a_R, b_R, i_L and j_L by a, b, i and j. The T_{ijab} and W_{ia} can be obtained by comparison with equation (5.2).

The partition function can be written as

$$Z = \sum_{V_{ia}} e^{-\beta\{\sum_{ijab} T_{ijab} V_{ia} V_{jb} + \sum_{ia} W_{ia} V_{ia}\}}$$

Observe that

$$\frac{-1}{\beta Z} \frac{\partial Z}{\partial W_{kc}} = \sum_{V_{ia}} V_{kc} \frac{e^{-\beta\{\sum_{ijab} T_{ijab} V_{ia} V_{jb} + \sum_{ia} W_{ia} V_{ia}\}}}{Z} = \bar{V}_{kc} \quad (5.3)$$

where $\bar{V}_{kc}$ is the mean of V_{kc} with respect to the probability distribution.

We can write the partition function in the form

$$Z = \sum_{V_{ia}} \prod_{ia} e^{-\beta V_{ia}\{\sum_{jb} T_{ijab} V_{jb} + W_{ia}\}}$$

The mean field approximation [119] consists of evaluating the partition function for each term V_{ia} while replacing the value of the other V_{jb}'s by their mean values $\bar{V}_{jb}$. This gives

$$Z \approx Z_{approx} = \prod_{ia} \{1 + e^{-\beta(\sum_{jb} T_{ijab} \bar{V}_{jb} + W_{ia})}\}$$

Observe that Z_{approx} is a function of the $\bar{V}_{jb}$'s, which are unknown. We can, however, use equation (5.3) to get consistency equations for the $\bar{V}_{ia}$. This gives us the mean field equations:

$$\bar{V}_{ia} = \frac{1}{1 + e^{+\beta\{\sum_{jb} T_{ijab} \bar{V}_{jb} + W_{ia}\}}} \quad (5.4)$$

To obtain a solution to these equations we use the deterministic update equation

$$\frac{d\bar{V}_{ia}(t)}{dt} = -\bar{V}_{ia}(t) + \frac{1}{1 + e^{+\beta\{\sum_{jb} T_{ijab} \bar{V}_{jb}(t) + W_{ia}\}}} \quad (5.5)$$

which can be proven to converge to a solution of (5.4). This update equation can also be shown to be a deterministic form of a standard Monte Carlo algorithm (see, for example, [101].

These update equations are an alternative to the update rule used in cooperative stereo algorithms ([37, 99]. In fact the update equations used in the cooperative stereo algorithms are equivalent to the limit of (5.5) as the constant $\beta \to \infty$. Since this limit corresponds to the low temperature limit of the system, where there are likely to be many local minima, we expect that the update rule in (5.5) run at finite β will be more effective than the cooperative stereo algorithm.

5.4 COMPARISONS WITH OTHER THEORIES

In the previous section we have used methods from statistical physics to analyze a general theory of stereo vision. In particular, we have shown how we can eliminate some of the fields to obtain equivalent, but superficially different, theories. In this section we show that this helps compare these theories to previous approaches.

5.4.1 THE MARR-POGGIO COOPERATIVE STEREO ALGORITHM

The operation of the Marr-Poggio cooperative stereo algorithms [99] involves determining the correspondence (or matching) between features in left and right images. Their matching system contains binary units C_{ia} for each lattice point ia (corresponding to a point i in the left image and a point a in the right image (note that, unlike our V_{ia}, these units do not occur only where there are features)).

The C_{ia} are initialized to be 1 if there are features at i and a in the left and right images, hence a potential match, otherwise they are 0.

They are updated by the following rule:

$$C_{ia}(t+\delta t) = \sigma\{k_1 \sum_E C_{jb}(t) - k_2 \sum_I C_{jb}(t) + k_3 C_{ia}(0)\}$$

where σ is a threshold function, E and I are excitatory and inhibitory neighborhoods, and the k_i are constants. This can be written as

$$C_{ia}(t+\delta t) = \sigma\{\sum_{jb} T_{ijab} C_{jb}(t) + k_3 C_{ia}(0)\}$$

where, for example, T_{ijab} may be of form

$$T_{ijab} = k_1\{\delta_{i,j-1}\delta_{a,b-1} + \delta_{i,j+1}\delta_{a,b+1}\}$$
$$-k_2\{\delta_{a,b}\delta_{i,j+1} + \delta_{a,b}\delta_{i,j-1} + \delta_{ij}\delta_{a,b+1} + \delta_{i,j}\delta_{a,b-1}\}$$

The algorithm corresponds to doing gradient descent with an energy function

$$E[C_{ia}] = \frac{1}{2}\sum_{i,j,a,b} T_{ijab} C_{ia} C_{jb} + \sum_{i,a} C_{ia}(0) C_{ia}$$

We now compare it to our theory without line process elements. This has energy function

$$E(V_{i_L a_R}) = \alpha \sum_{i,j}(\sum_a V_{ia}(x_i - x_a))(\alpha\delta_{ij} + G(x_i, x_j))^{-1}$$
$$\times(\sum_b V_{jb}(x_j - x_b)) + \alpha\sum_{i_L}(\sum_{a_R} V_{i_L a_R} - 1)^2$$
$$+\alpha\sum_{a_R}(\sum_{i_L} V_{i_L a_R} - 1)^2$$

Comparing the two energy functions we see a number of similarities. There is a general excitation for a match in the direction of constant disparity, due to the first term, and inhibition of matching in the viewing directions, due to the second term. The principle difference is that our method only has matching elements at feature points, rather than everywhere in the image. This may result in a difference between the systems for transparent surfaces [171].

This suggests an alternative strategy for minimizing the energy function for the first level theory by performing steepest descent in the energy function as for cooperative stereo. This is, as mentioned in Section 5.2.1, a special case of the mean field method in the limit as $\beta \to \infty$.

By the mathematical connection to the traveling salesman problem given in [166] and the empirical comparison of different algorithms for that problem we believe that the cooperative stereo algorithms will be less successful than the level 1 theory using disparity fields only, see Section 5.1.2.

5.4.2 DISPARITY GRADIENT LIMIT THEORIES

It is interesting to contrast the second level theory with theories matching features based on the disparity gradient limit (Pollard, Mayhew and Frisby, [121]; Prazdny, [129]). In Prazdny's algorithm a possible match between two points is supported by all possible matches with similar disparities. Matches with very different disparities give negligible support for the match, but do not inhibit it. This means the theory does not assume a surface with smoothly varying disparity and enables the theory to deal with transparent surfaces. In our notation this theory can be formalized in terms of *maximizing* a cost function $E(V_{ai})$ over all matching assignments V_{ai}

$$E(V_{i_L a_R}) = \sum_{i_L, j_L} e^{-\frac{\left(\sum_{a_R}(x_{i_L}-x_{a_R})V_{i_L a_R} - \sum_{b_R}(x_{j_L}-x_{b_R})V_{j_L b_R}\right)^2}{2\sigma^2 |x_{i_L}-x_{j_L}|^2}}$$

Here $\sum_{a_R}(x_{i_L} - x_{a_R})V_{i_L a_R}$ corresponds to the disparity at point x_{i_L}. Thus a potential match of point i_L to point a_R is supported by the other matches. The support is a Gaussian function depending on the difference between the two disparities divided by the distance between the points on the image.

There is some similarity to our second level theory. After we have averaged out the discontinuity fields the theory is described by an effective energy

$$\begin{aligned} E_{eff}(d,V) &= \textstyle\sum_k \gamma - \frac{1}{\beta}\log(1 + e^{-\beta[\alpha(d_k - d_{k-1})^2 - \gamma]}) \\ &+ \textstyle\sum_{a_R, i_L} V_{i_L a_R}\Big\{[d(x_{i_L}) - (x_{a_R} - x_{i_L})]^2 + \lambda[\sum_{b_R} V_{i_L a_R} V_{i_L b_R} \\ &+ \textstyle\sum_{j_L} V_{i_L a_R} V_{j_L a_R} - 2V_{i_L a_R}]\Big\} \end{aligned} \quad (5.6)$$

This theory matches points so as to obtain the smoothest possible disparity field except at discontinuities. These discontinuities occur at the threshold $|d_k - d_{k+1}| > \sqrt{(\gamma/\alpha)}$ (see equation (5.6), this threshold is analyzed for the image segmentation case by Geiger and Girosi [46], and it is exact in the limit as $\beta \to \infty$). This threshold is roughly equivalent to the limit used in the disparity gradient limit theories and performs the same function.

The P.M.F. theory (Pollard, Mayhew and Frisby [121]) used a value for the disparity gradient limit based on experimental results by Burt and Julesz [27] using dot stimuli. The experiments of Bülthoff and Fahle [24] suggest that this limit is a function of the stimulus and seem more consistent with our theory in which the parameters α and γ are determined by the stimulus.

Once again, the second level theory described in terms of the effective energy for the disparity field, as given in Section 5.1.4, should be better than theories which include matching elements.

5.5 COMPARISONS WITH PSYCHOPHYSICAL DATA

In this section we briefly describe the relationship between our theory and psychophysics (more details of the comparison are provided in [169]. We will chiefly be concerned with two experiments [25, 26, 24] which provide measurements of the perceived depth as a function of the real depth and the matching primitives by use of reference systems (or depth probes). These experiments are particularly useful for our purposes because of: (i) the quantitative depth information they supply and (ii) their investigation of which features are used for matching and how the perceived depth depends on these features.

One overall conclusion from these experiments is that objects perceived stereoscopically tend to be biased towards the fronto-parallel plane and the degree of this bias depends on the features being matched. This is in general agreement with our theory in which the disparity smoothness term causes such a bias with a magnitude depending on the robustness and discriminability of the features.

For the first set of experiments [24] the observer was asked to estimate the disparity of stereo stimuli relative to a set of reference lines. The stimuli were either lines at various angles or pairs of dots or features.

The experiments showed that the perceived disparity decreased as a function of the disparity gradient. This effect was: (a) strongest for horizontal lines, (b) strong for pairs of dots or similar features, (c) weak for dissimilar features and (d) non-horizontal lines.

Our explanation assumes these effects are due to the matching strategy and is based on the second level theory, with energy function given by (5.2). The idea is that the smoothness term (the third term) is required to give unique matching but that its importance, measured by γ, increases as the features become more similar. If the features are sufficiently different (perhaps pre-attentively discriminable) then there is no matching ambiguity, so the correct disparities are obtained. If the features are similar then smoothness (or some other a priori assumption) must be used to obtain a unique match, leading to biases towards the fronto-parallel plane. The greater the similarity between features the more the need for smoothness and hence the stronger the bias towards the fronto-parallel plane. The discontinuity field is switched on at both the points ensuring that smoothness is only imposed between the two points. Thus the two points are considered the boundaries of an object and only the object itself is smoothed.

An analysis of the second level theory [169] shows that it predicts the falloff of perceived disparity with disparity gradient, provided we choose the smoothness operator to be the first derivative of the disparity. The change of rate of falloff for different types of features is due to varying γ as described above.

The results are not consistent with several possible choices of the smoothness operator, such as the second derivative of the disparity $\partial^2 d/\partial x^2$. It is straightforward to calculate that this choice does not bias towards the fronto-parallel plane. It is likely that the smoothness operator S must contain a $\partial/\partial x$ term to ensure the observed fronto-parallel bias.

The second experiments [25, 26] compared the relative effectiveness of image intensity and edges as matching primitives. The stimuli were chosen to give a three dimensional perception of an ellipsoid. The observer used a stereo depth probe to make a pointwise estimate of the perceived shape.

The experiments showed that depth could be derived from images with disparate shading even in the absence of disparate edges. The perceived depth, however, was weaker for shading disparities (seventy percent of the true depth).

Putting in edges or features helped improve the accuracy of the depth perception. But in some cases these additional features appeared to decouple from the intensity and were perceived to lie above the depth surface generated from the intensity disparities.

These results are again in general agreement with our model. The edges give good estimates of disparity and so little *a priori* smoothness is required and an accurate perception results. The disparity estimates from the intensity, however, are far less reliable (small fluctuations of intensity might yield large fluctuations in the disparity). Therefore more *a priori* smoothness is required to obtain a stable result. This gives rise to a weaker perception of depth.

The use of the image intensity peak as a matching feature is vital (at least for the edgeless case) since it ensures that the image intensity is accurately matched. For these images, however, the peak is difficult to localize and depth estimates based on it are not very reliable. Thus the peak is not able to pull the rest of the surface to the true depth.

Bülthoff and Mallot [26] found that pulling up did occur for the edgeless case if a dot was added at the peaks of the images. This is consistent with our theory since, unlike the peaks, the dots are easily

localized and matching them would give a good depth estimate. Our present theory, however, is not consistent with a perception that sometimes occurred for this stimulus. In some cases the dots were perceived as lying above the surface rather than being part of it. This may be explained by the extension of the stereo theory described in this chapter to transparent surfaces [171].

5.6 CHAPTER SUMMARY

- We described the theory of stereo vision proposed in [169]. This theory employs the Bayesian approach to vision, and uses the techniques described in chapter 3. This theory is able to incorporate most of the desirable elements of stereo and it is closely related to a number of existing theories of stereo vision.

- The stereo vision algorithms can combine information from matching different primitives, which is desirable on computational and psychophysical grounds. The formulation can be extended to include monocular depth cues for stereo correspondence, as is shown in the next chapter.

- A basic assumption of the stereo algorithms described here is that correspondence and interpolation should be performed simultaneously. This is related to the important experimental and theoretical work of Mitchison [111] and Mitchison and McKee [112].

- The use of mean field theory enables us to average out, in various combinations, the disparity and matching fields, enabling us to make mathematical connections between different formulations of stereo. It also suggests novel algorithms for computing the estimators (due to enforcing the matching constraints while performing the averaging) and we argue that these algorithms are likely to be more effective than a number of existing algorithms.

- Finally the theory agrees well with some psychophysical experiments [25, 26, 24], although further experiments to investigate the importance of different stereo cues are needed.

Chapter 6

Fusing Binocular and Monocular Depth Cues

6.1 STRONG FUSION – STEREO WITH MONOCULAR CUES

In this chapter we develop a technique for strong coupling of modules. As explained in chapter 4, strong coupling occurs when one module interacts during the computation of another. For example if the modules are stereo and eye-movements, then the eye-movements directly affect the matching of the stereo process. We will primarily concentrate on methods for strongly coupling stereo with monocular cues. The formalism, however, is more general and can be directly adapted to other matching problems such as long range motion correspondence. It will be shown that the theory for stereo integration of different features that described in the previous chapter can be naturally extended to incorporate these additional cues. Again methods from statistical mechanics, or statistics, can be applied to impose the matching constraints in a "strong" manner.

We will begin by briefly describing two previous attempts at strong coupling. Geiger and Yuille [48] described a method for combining stereo with eye and head movements while House [67] constructed a method for combining stereo and accommodation cues.

6.2 PREVIOUS ATTEMPTS AT STRONG COUPLING FOR STEREO

A key problem in stereo vision is to determine the correspondence between features in the two eyes. In chapter 5 we described how heuristic *a priori* assumptions were used to determine a unique solution. These assumptions, such as smoothness with/without discontinuities, are suitable for many real world images but will inevitably fail in some situations. It is sensible to use other additional information to decrease the reliance on these heuristic constraints. This additional information might be high level, supplied by domain dependent knowledge or by an object recognition process, or it might be low level. In this book we will restrict ourselves to low level information and, in particular, to monocular cues.

In [48] there is proposed a method that uses monocular cues determined by eye movements (or head movements) to supply the additional information. By rotating the eyes to alter the direction of fixation we introduce extra views of the object (this assumes that the center of rotation does not lie of the image plane, otherwise no new information could be gained from these additional views). This corresponds to having several views of the object and is similar to doing stereo with three or more cameras [117]. Since the eye movements are small it is possible to track features in each eye as the eye rotates. Using triangulation we can then obtain monocular depth estimates for the features.

Since eye movements are small the monocular depth information they yield is very unreliable and for this reason it is commonly

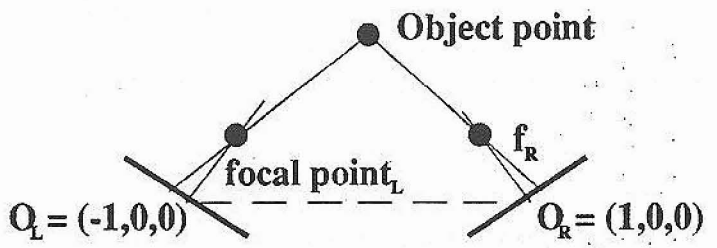


Figure 6.1: Imaging geometry.

thought that they only supply weak information. However it was shown in [48] that indeed they could supply a lot of information provided the sources of error involved are taken into account. They assumed that these errors arose from lattice quantization (refinements to the theory would take into account errors of localization in the feature detection process). In their theory the two eyes provided error estimates in addition to monocular depth estimates. These error estimates were typically very large but they could still be used to help solve the stereo correspondence.

Assume the imaging geometry, shown in figure 6.1, used by Longuet-Higgins [90]. A point $\vec{X} = (x, y, z)$ in space is projected to points $P_r(\phi)$ and $P_l(\psi)$ in the right and left eyes. The coordinates of these points are given by

$$x_r(\phi) = -f\frac{(x-l)\cos\phi - y\sin\phi}{(x-l)\sin\phi + y\cos\phi - f}$$

$$z_r(\phi) = -f\frac{z}{(x-l)\sin\phi + y\cos\phi - f}$$

$$x_l(\psi) = -f\frac{(x+l)\cos\psi - y\sin\psi}{(x-l)\sin\psi + y\cos\psi - f}$$

$$z_l(\psi) = -f\frac{z}{(x+l)\sin\psi + y\cos\psi - f} \tag{6.1}$$

If we consider the left eye, as we vary ϕ we get a set of measurements $(x_r(\phi), z_r(\phi))$ from which we can, in principle, calculate (x, y, z). From two points $(x_r(\phi_1), z_r(\phi_1))$ and $(x_r(\phi_2), z_r(\phi_2))$ we calculate

$$\begin{aligned} x - l &= \frac{B(x_r(\phi_1), z_r(\phi_1), x_r(\phi_2), z_r(\phi_2))}{A(x_r(\phi_1), z_r(\phi_1), x_r(\phi_2), z_r(\phi_2))} \\ y &= \frac{C(x_r(\phi_1), z_r(\phi_1), x_r(\phi_2), z_r(\phi_2))}{A(x_r(\phi_1), z_r(\phi_1), x_r(\phi_2), z_r(\phi_2))}, \\ z &= -\frac{z_r(\phi_1)D_1}{f} \end{aligned} \tag{6.2}$$

where A, B, and C are quadratic functions of $x_r(\phi_1)$, $y_r(\phi_1)$, $x_r(\phi_2)$, and $y_r(\phi_2)$) and D_1 depends only on ϕ_1. By differentiating these expressions we obtain a linear relation between the errors $\delta x, \delta y, \delta z$ of the spatial position (x, y, z) in terms of the errors $\delta_1, \delta_{z1}, \delta_2, \delta_{z2}$ in the measurements of $(x_r(\phi_1), y_r(\phi_1), x_r(\phi_2), y_r(\phi_2))$ (computer experiments show that this the second order terms are negligible and can be ignored). We now assume that the δ's are distributed independently in the range $(-L/2, L/2)$ where L is the lattice spacing. This induces a probability distribution on the errors $\delta x, \delta y \delta z$ from which we can calculate the means and the standard deviations, hence giving us an estimate of the reliability of the monocular estimates.

The proposed strategy falls into two parts. First we track (and match) features in the left and right eyes separately. Each eye gives a rough estimate for the 3-D position of the point and the error range for this position (a multiple of the standard deviation). We now use these estimates as the basis for the stereo match. Second we define a *rotation depth test* which accepts a possible match if the estimated positions and the error range are compatible and some other, more technical, tests. The strictness of these tests depends on a set of *control parameters*. Initially these parameters are set to make the tests very difficult. The program now hypothesizes matches between points in the right and left eyes. If the tests are all passed then these matches are accepted and the points are not considered further. Otherwise the points are left unmatched. Then the algorithm changes the control parameters to reduce the strictness of the test and matches are again hypothesized and tested. This procedure is repeated until the control parameters reach a final value (determined by the amount of error accepted by the system). Points which have not been matched by the algorithm, perhaps because they occur at occluding boundaries and are only visible to one eye, are assigned the 3-D position estimated by the rotation of the eyes. A "zooming" feature can be added to the basic algorithm: if a certain region of the image contains a large number of points then the eyes can zoom in to this region by changing the focal length, thereby increasing the resolution.

The strategy essentially matches the points whose probability distributions overlap above a certain threshold, removes these points thereby reducing the ambiguity for the remaining points, lowers the threshold and repeats the process.

More precisely: suppose we have two points P_R and P_L in the right and left eyes which are possible matches with monocular position estimates of (x_R, y_R, z_R) and (x_L, y_L, z_L) with standard deviations $(\sigma_{xR}, \sigma_{yR}, \sigma_{zR})$ and $(\sigma_{xL}, \sigma_{yL}, \sigma_{zL})$. The rotation depth test is specified by three parameters C_1, C_2, C_3). Then P_R and P_L will pass the test provided

$$\begin{aligned} |x_L - x_R| &\leq C_1(\sigma_{xR} + \sigma_{xL}) \\ |y_L - y_R| &\leq C_2(\sigma_{yR} + \sigma_{yL}) \\ |z_L - z_R| &\leq C_3(\sigma_{zR} + \sigma_{zL}) \end{aligned}$$

Additional tests are also used. For example the *stereo test* checks consistency between the monocular depth estimates of two possibly matching points and the stereo depth (and error) that would be assigned if the points did indeed match.

This algorithm was implemented and performed well on a number of synthetic images on which traditional stereo algorithms using

smoothness constraints would fail. Although it was somewhat ad hoc and did not provide an overall framework for strong coupling it nevertheless make use of a number of key ideas, namely the importance of estimating the errors of the monocular cues and the strategy of matching the most probable features first.

Another algorithm for strong coupling was developed by House [67] who sought to integrate stereo and accommodation depth cues. This work was strongly influenced by, and attempted to model, experiments by Collett and coworkers [38] on how frogs and toads use these visual cues.

The model is expressed in terms of a cooperative algorithm similar to that of Dev [37] and Marr and Poggio [99]. The monocular accommodation cues were used as an additional modification to the stereo algorithm. Although this algorithm was successful there are two criticisms we can make: (i) there was no estimate for the errors in the accommodation or stereo process, and (ii) the constraints for unique matches were imposed weakly as in the cooperative stereo algorithm, rather than by a strong method (see the previous chapter).

In the next section we will propose a general formalism for strong coupling of stereo with monocular cues. If the constraints are imposed in a weak manner this gives a theory closely related to House's work. The analysis is an extension of the argument in section 5.4 which showed that cooperative stereo theories are related to the Yuille, Geiger, Bülthoff [169] stereo theory with the constraints imposed weakly.

6.2.1 A GENERAL FRAMEWORK

Suppose we have a set of monocular measurements $(x_i^l, d_i^l, \sigma_i^l)$ and $(x_a^r, d_a^r, \sigma_a^r)$ and a depth from stereo function $d_s(x_i^l, x_a^r)$ with standard deviation $\sigma_s(i,a)$ (which is a measure of the uncertainty of the value of the depth obtained with the stereo module). Here x_i^l is a point

in the left image with a depth estimate d_i^l, given by eye-movements or focusing or some other monocular cue, and σ_i^l is the standard deviation of this estimate. Similarly, x_a^r is a point in the right image (a possible match to x_i^l) that also has a depth estimate, d_a^r, associated with it obtained from some monocular (single view) method, having a standard deviation given by σ_a^r.

Suppose we know that points i corresponds to points a_i. The best mean squared estimate of the depth $d(x_i)$, assuming Gaussian errors, is obtained by minimizing

$$E(f) = \sum_i \frac{1}{\sigma_i^l}(f(x_i^l) - d_i^l)^2 + \sum_i \frac{1}{\sigma_{a_i}^r}(f(x_i^l) - d_{a_i}^r)^2 + \sum_i \frac{1}{\sigma_{s(i,a_i)}}(f(x_i^l) - d_s(i,a_i))^2 \quad (6.3)$$

This approach corresponds to a weakly coupled method since the depth from eye movements and depth from stereo modules operate independently of each other. In most cases we do not know the correspondences and we want a method of finding it. One possibility is to define binary matching elements V_{ai} as before and an energy function

$$E(V_{ai}, f) = \sum_i \frac{1}{\sigma_i^l}(f(x_i^l) - d_i^l)^2 + \sum_{a,i} \frac{1}{\sigma_a^r} V_{ai}(f(x_i^l) - d_a^r)^2 + \sum_{a,i} \frac{1}{\sigma_s} V_{ai}(d(x_i^l) - f_s(i,a))^2 \quad (6.4)$$

We set $V_{ai} = 1$ if point a in one eye is matched to point i in the other eye, otherwise $V_{ai} = 0$. Note that if the correct or optimal matching is known or given a priori this cost function reduces to the previous cost function. This energy function should be minimized over the set of all possible matches V_{ai}, with the constraint that all points usually have a unique match.

Once again we can also include a priori terms in the energy function, typically smoothness terms. These terms are not needed to

give the energy function a unique minimum, but they are required to give dense depth values. We could, for example, include a term

$$E_{smooth}(f) = \gamma \sum (f_i - f_{i+1})^2 \tag{6.5}$$

where γ is a constant. This would correspond to a purely smoothing term. We could also modify it, as in the stereo chapter, to include discontinuity fields to break the smoothness constraints.

6.2.2 SOFT AND HARD CONSTRAINTS

There are two main approaches for minimizing $E(V_{ai}, f)$ depending on whether we impose the global conditions on $\{V_{ai}\}$ by biases in the energy function (soft constraints) or by some implicit manner (hard constraints). As we argued in the stereo chapter hard constraints seem strongly preferable to soft constraints.

Our analysis now closely follows our work in the stereo chapter. We have a cost function $E(V_{ai}, f)$ depending on two fields and we can choose to eliminate either field.

We can eliminate the f field directly (since $E(V_{ai}, f)$ is quadratic in f). As in section (stereo) we will obtain an effective energy $E_{eff}(V_{ai})$ depending only on the $\{V_{ai}\}$. The uniqueness constraints on the matching can now be imposed by adding a constraint term

$$E_{constraint}(V_{ai}) = \sum_a (\sum_i V_{ai} - 1)^2 + \sum_i (\sum_a V_{ai} - 1)^2 \tag{6.6}$$

This yields a cost function $E_{eff}(V_{ai}) + E_{constraint}(V_{ai})$ which is quadratic in the V_{ai}. We can directly turn this into a cooperative algorithm by doing iterated steepest descent of this energy function. Alternatively we can introduce a temperature T and then write a system of continuous differential equations converging to the mean field equation (see chapter 5). These equations are very close to those used by House [67] for coupling stereo with accommodation.

We prefer, however, the alternative approach (see stereo section) in which the matching fields $\{V_{ai}\}$ are eliminated by summing over a restricted class of $\{V_{ai}\}$ which satisfy (mostly) the global constraints. In this case it is sufficient to sum over states in which each points in the left and right eyes have a unique match (without needing to require that these matches are the same).

This gives an effective energy

$$\begin{aligned} E_{eff}(f) = & \sum_i \frac{1}{\sigma_i^l}(f(x_i^l) - d_i^l)^2 \\ & -\frac{1}{\beta}\sum a \log\left[\sum_i \exp\{-\beta\frac{(f(x_i^l) - d_a^r)^2}{2\sigma_a^r}\}\right] \\ & -\frac{1}{\beta}\sum_i \log\left[\sum_a \exp\{-\beta\frac{(f(x_i^l) - d_a^r)^2}{2\sigma_a^r}\}\right] \\ & -\frac{1}{\beta}\sum_a \log\left[\sum_i \exp\{-\beta\frac{(f(x_i^l) - d_s(i,a))^2}{2\sigma_s}\}\right] \\ & -\frac{1}{\beta}\sum_i \log\left[\sum_a \exp\{-\beta\frac{(f(x_i^l) - d_s(i,a))^2}{2\sigma_s}\}\right] \end{aligned} \tag{6.7}$$

where β is the inverse temperature of the system.

We can now minimize this cost function be steepest descent starting with large temperature (small β) and then gradually reducing it. This is somewhat analogous to the method described in [48] since using small β corresponds to small control parameters and hence makes the depth tests hard to satisfy. However it differs by not removing features once they have been matched at small β thereby reducing the ambiguity for other matches. But we expect that this will be taken into account by hysteresis.

We can add additional *a priori* constraints to the cost function if required. For example, it is straightforward to adapt the smoothness and discontinuity constraints from the stereo chapter. In general it is better not to include such constraints unless the monocular depth data is very unreliable.

It is hard to estimate the reliability of some monocular cues. Shape from defocus is one such example [120, 128, 69, 19, 145]. The methods either involve strong assumptions about the "true" image, for example that it is an ideal step edge, or (often implicit) assumptions about the local smoothness of the surface. To incorporate these cues into strongly coupled shape modules would require a systematic analysis of the assumptions and the errors involved. A start to such an analysis has been attempted by Hwang and coworkers [69], and we see no conceptual difficulty in using it as an input into our framework for strong coupling.

We may also assign a measure of uncertainty to the result of the process. We could then propagate the best possible set of matches to future processes. A more attractive alternative is to assign each set of possible depths f a probability proportional to $exp(-E_{eff}[f]/T)$. The factor T would act as a coarse-fine parameter. If T is small only one state will be likely. If T is large there will be more possibilities.

6.3 SUMMARY

- Monocular cues can be used to disambiguate correspondence problems even if the cues are unreliable, provided we can estimate this unreliability.
- It is straightforward to generalize the stereo framework to incorporate strong coupling for matching problems. Techniques from statistical physics can be applied as before to generate algorithms for solving these problems.
- For some monocular cues, such as depth from defocus, we need detailed modeling of the process to determine the assumptions used and the errors involved before the monocular cues can be used for strong coupling.

Chapter 7

Data Fusion in Shape From Shading Algorithms

The vision process that we examine in this chapter is that of obtaining object shape from information about the specular and Lambertian components of the light reflected off of the object. We demonstrate the application of two different forms of weakly coupled fusion to the problem of estimating object shape from shading information. The first example is of an algebraic approach, while the second involves an energy function or Bayesian approach. We follow this description of the weakly coupled methods with an illustration of how we can convert the weakly coupled fusion algorithms into a strongly coupled fusion algorithm.

In each of the shape from shading algorithms to be described we will assume that we have available two spatially and temporally registered images, corresponding to the specular and Lambertian components of the surface reflectance. It is possible to extract the specular and Lambertian reflectance components when we have an image containing multiple wavelength bands (i.e. we have available a color image of the scene). If we assume a dichromatic model of surface reflectance [86, 80], it is possible to use color information to distinguish

between these two reflectance components [53, 86, 139]. This is due to the fact that, in the dichromatic model, the color of the Lambertian component is that of the object surface, while the color of the specular component is that of the illumination. If the illuminant and object body colors are different then the colors of the points in the image will typically lie on two non-collinear lines in color space (i.e. Red-Green-Blue space). One of these lines in the color space will correspond to the specular image component, while the other will correspond to the Lambertian image component. In this manner the color image can be converted into a pair of images, one containing the amount of specular reflectance, the other containing the level of Lambertian reflectance.

Another approach, due to Wolff [160] for extracting the specular and Lambertian reflectance components involves the use of polarizing filters. The basic idea behind this technique is that the polarization of light reflecting off of a surface is different for specular reflection than it is for diffuse (or Lambertian) reflection. In particular, the diffuse reflections are generally unpolarized while the specular reflections are biased towards the normal to the specular plane of incidence. If one plots the orthogonally polarized image components $k_\perp$ and $k_\parallel$ (which are obtained by viewing through polarizing filters) as coordinates in a 2D space, it can be shown that the image points near a specularity will map into a line segment. The specular and diffuse reflectance values can be obtained from the slope and intercept of this line segment [160].

7.1 AN ALGEBRAIC APPROACH TO FUSING SPECULAR AND LAMBERTIAN REFLECTANCE DATA

However the specular and Lambertian surface reflectance components are obtained, they can directly provide surface normal values,

assuming the following simple reflectance models. A simple model for the reflectance function of specularities is given by

$$E_s(x,y) = (\hat{k} \cdot \hat{h}(x,y))^m = R_s(\hat{n}) \tag{7.1}$$

where $E_s(x,y)$ is the specular image intensity (normalized to lie between 0 and 1), $\hat{k}$ is a unit vector in the viewer direction, $\hat{s}$ is a unit vector in the light source direction, $\hat{n}$ is the surface unit normal vector and $\hat{h}$ is a unit vector that depends on $\hat{s}$ and $\hat{n}$

$$\hat{h}(x,y) = 2(\hat{n}(x,y) \cdot \hat{s})\hat{n}(x,y) - \hat{s} \tag{7.2}$$

The parameter m is a number (assumed to be known) corresponding to the sharpness of the specularity. To obtain surface shape information from the specular reflectance we must invert equation (1) to get

$$(\hat{k} \cdot \hat{h}(x,y)) = E_s(x,y)^{1/m} \tag{7.3}$$

We also have the Lambertian shading component $E_l(\hat{n})$ given by the Lambertian model

$$E_l(x,y) = (\hat{n}(x,y) \cdot \hat{s}) = R_l(\hat{n}) \tag{7.4}$$

The functions $R_s(\hat{n})$ and $R_l(\hat{n})$ are the *reflectance maps* for specular and Lambertian surfaces respectively. Each of the image reflectance components (i.e. the specular or Lambertian components) is, by itself, insufficient for the task of obtaining a unique solution to the shape extraction problem, as in each case we have only one equation for the two unknowns (the x and y components of the surface unit normal vectors).

Using either the color based or polarization based methods of extracting specular and Lambertian reflectance described earlier we can obtain both E_s and E_l. We therefore now have two equations for the surface normal, one from each reflectance component, and so theoretically there is enough information to determine the two components of the surface normal uniquely.

Combining the equations for the surface normal vectors as a function of the specular and Lambertian components we can see that

(after some algebra):

$$\hat{n}(\hat{x}) = \alpha(\hat{x})\hat{s} + \beta(\hat{x})\hat{k} + \gamma(\hat{x})\hat{s} \times \hat{k} \tag{7.5}$$

where

$$\alpha = \frac{1}{s^2\Omega}\left(E_l(\vec{x}) - \frac{E_s^{\frac{1}{m}}(\hat{x})c\Omega}{2E_l(\hat{x})} - \frac{c^2\Omega}{2E_l(\hat{x})}\right) \tag{7.6}$$

$$\beta = \frac{1}{s^2\Omega}\left(\frac{E_s^{\frac{1}{m}}(\vec{x})}{2E_l(\hat{x})} + \frac{c\Omega}{2E_l(\hat{x})} - c\Omega E_l(\vec{x})\right) \tag{7.7}$$

$$\gamma = \pm\frac{(4s^2\Omega E_l^2 - 4E_l^4 - (E_s^{\frac{1}{m}} + c\Omega)^2 + 4E_l^2(E_s^{\frac{1}{m}} + c\Omega)c\Omega)^{1/2}}{2E_l s\Omega} \tag{7.8}$$

with $c\Omega = \cos\Omega = \hat{k} \cdot \hat{s}$ being the cosine of the angle between the viewer direction and the light source, and $s\Omega = \sin\Omega$. Note that there is an ambiguity in the sign of γ, which will manifest in an ambiguity in the sign of the component of the unit normal vector in the direction of $\hat{s} \times \hat{k}$. This ambiguity *cannot* be resolved using only the specular and Lambertian image components, since both $R_s(\hat{n})$ and $R_l(\hat{n})$ are symmetrical about $\hat{n} = \hat{s} \times \hat{k}$. Thus, in order to obtain a unique surface normal, we must provide more information than just E_s and E_l. Typically this will be in the form of a smoothness constraint combined with boundary conditions (such as provided by the occluding contour of the surface). However, such a constraint is difficult to embed in an algebraic algorithm. Hence, in our examination of the algebraic approach we will assume that the information regarding the sign of γ is provided by an independent module, which we will leave unspecified.

In the above algorithm we have fused two data sources, $E_s(\hat{x})$ and $E_l(\hat{x})$, and a source of *a priori* knowledge (concerning the sign of γ) to provide the object shape, parametrized by surface normals $\vec{n}(\hat{x})$. Following the classification of data fusion algorithms put forth in chapter 4, this is a class II weakly coupled fusional process, as no relative weighting of the two data sources is performed and there is no coupling between the shape from shading module back to the reflectance component extraction process. This method is similar in

spirit to the active vision algorithms developed by Aloimonos et al [2, 1] in that it uses two independent sources of data to algebraically solve a system with two unknowns (namely the components of the normal vector at a point on a surface).

To investigate the uncertainty of the computed shape we see how small changes in the input data affect the output. Suppose we have perturbations δE_l and δE_s in the input data. We then get changes in the output of the fusional module given by

$$\delta\hat{n} = (\delta\alpha)\hat{s} + (\delta\beta)\hat{k} + (\delta\gamma)\hat{s} \times \hat{k} \tag{7.9}$$

where

$$\delta\alpha = \frac{1}{s^2\Omega}\left(\left[1 + \frac{E_s^{1/m}c\Omega}{2E_l^2} + \frac{c^2\Omega}{2E_l^2}\right]\delta E_l - \left[\frac{c\Omega}{2mE_lE_s^{(m-1)/m}}\right]\delta E_s\right) \tag{7.10}$$

$$\delta\beta = \frac{1}{s^2\Omega}\left(-\left[c\Omega + \frac{c\Omega}{2E_l^2} + \frac{E_s^{1/m}}{2E_l^2}\right]\delta E_l + \left[\frac{1}{2mE_lE_s^{(m-1)/m}}\right]\delta E_s\right) \tag{7.11}$$

$$\begin{aligned}\delta\gamma = &\frac{-1}{\tau s^2\Omega}\left[\frac{(E_s^{1/m} + c\Omega)^2 - 4E_l^4}{2E_l^2}\right]\delta E_l \\ &+\frac{-1}{\tau s^2\Omega}\left[\frac{2c\Omega E_l^2 - E_s^{1/m} - c\Omega}{2mE_lE_s^{(m-1)/m}}\right]\delta E_s\end{aligned} \tag{7.12}$$

where

$$\tau = \pm(4s^2\Omega E_l^2 - 4E_l^4 - (E_s^{1/m} + c\Omega)^2 + 4E_l^2(E_s^{1/m} + c\Omega)c\Omega)^{1/2} \tag{7.13}$$

It is clear from these equations that there will be instabilities as $\Omega \mapsto 0$, $E_l \mapsto 0$, $E_s \mapsto 0$ and as $\tau \mapsto 0$. Since the instability as $\tau \mapsto 0$ occurs when $\hat{s}$, $\hat{n}$ and $\hat{k}$ are coplanar and $\gamma = 0$, it is non-generic in the sense that any small perturbation of $\hat{n}$ will move one away from the locus of instability. The instability as $\Omega \mapsto 0$, however, cannot be avoided as this case occurs when $\hat{k}$ and $\hat{s}$ are parallel and occurs independently of the value of $\hat{n}$. In this situation information about

only one component of $\hat{n}$ is available. Similarly the instabilities as $E_l \mapsto 0$ and $E_s \mapsto 0$ cannot be avoided.

If one can obtain the probability distributions of the uncertainties, δE_l and δE_s, of the input data, then we can, in principle, derive expressions for the probability distribution of $\delta\hat{n}$ and compute its variance and other statistical properties. Thus we can produce a measure for the uncertainty of the fused information which can be used in subsequent fusional processes. In general, however, the derivation of the probability density of $\delta\hat{n}(\hat{x})$ will be exceedingly difficult, if not impossible, due to the nonlinearity of equation (7.9). Even without going through the exercise of deriving these probability densities we can observe that the variance of $\delta\hat{n}(\hat{x})$ will depend on the measured values of E_l and E_s. It will be small if the values of E_s and E_l are large and becomes infinite as either tends to zero. A simpler, alternative, approach to the generation of uncertainty measures is to look at the product of the uncertainty in the inputs with the sensitivity of the computed surface normal to changes in the inputs. That is, we linearize the variation of $\delta\hat{n}$ with respect to δE_s and δE_l to give:

$$\delta\hat{n} = \left(\frac{\partial\hat{n}(E_s, E_l, \hat{s}, \hat{k})}{\partial E_s}\right)\delta E_s + \left(\frac{\partial\hat{n}(E_s, E_l, \hat{s}, \hat{k})}{\partial E_l}\right)\delta E_l \qquad (7.14)$$

where $\frac{\partial\hat{n}}{\partial E_s}$ and $\frac{\partial\hat{n}}{\partial E_l}$ can be determined from equation (7.5) and the equations for $\delta\alpha, \delta\beta$, and $\delta\gamma$ to give:

$$\frac{\partial\hat{n}}{\partial E_s} = \frac{\left[\hat{k} - \hat{s}c\Omega + \frac{1}{\tau}(\hat{s}\times\hat{k})(2c\Omega E_l^2 - E_s^{\frac{1}{m}} - c\Omega)\right]}{2mE_lE_s^{\frac{m-1}{m}}s\Omega} \qquad (7.15)$$

$$\frac{\partial\hat{n}}{\partial E_l} = \frac{\hat{s}(2E_l^2 + E_s^{\frac{1}{m}}c\Omega + c^2\Omega) - \hat{k}(2E_l^2c\Omega + c\Omega + E_s^{\frac{1}{m}})}{2E_l^2s\Omega} - \frac{(\hat{s}\times\hat{k})((E_s^{\frac{1}{m}} + c\Omega)^2 - 4E_l^4)}{2\tau E_l^2 s\Omega} \qquad (7.16)$$

If we assume that δE_s and δE_l are uncorrelated zero mean Gaussian white noise processes with variances σ_s^2 and σ_l^2 respectively, then the

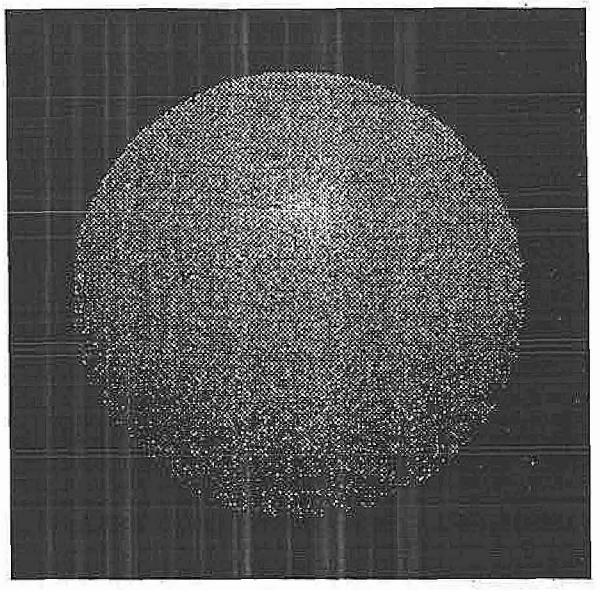


Figure 7.1: The original noisy view of a Lambertian shaded sphere with specularity.

linear approximation to $\delta\hat{n}$ given above will be zero mean Gaussian as well. We can, therefore compute the variance of (x and y components of) $\delta\hat{n}(\hat{x})$ as:

$$\hat{\sigma}_n^2 = E\{\delta\hat{n}^2\} = E\{(\delta n_x^2, \delta n_y^2)\} = \left(\frac{\partial\hat{n}}{\partial E_s}\right)^2\sigma_s^2 + \left(\frac{\partial\hat{n}}{\partial E_l}\right)^2\sigma_l^2 \qquad (7.17)$$

where σ_s^2 and σ_l^2 are the variances of the noise in the specular and Lambertian image components, respectively. In the above derivation we used the fact that the expected value of $\delta E_s\delta E_l$ is zero since we assume that δE_s and δE_l are uncorrelated.

Our algorithm was run on a synthetic image of a sphere created using the specular and Lambertian reflectance models of equations (1) and (4), where it was assumed that the extraction of the specular and Lambertian components, perhaps by a color based scheme,

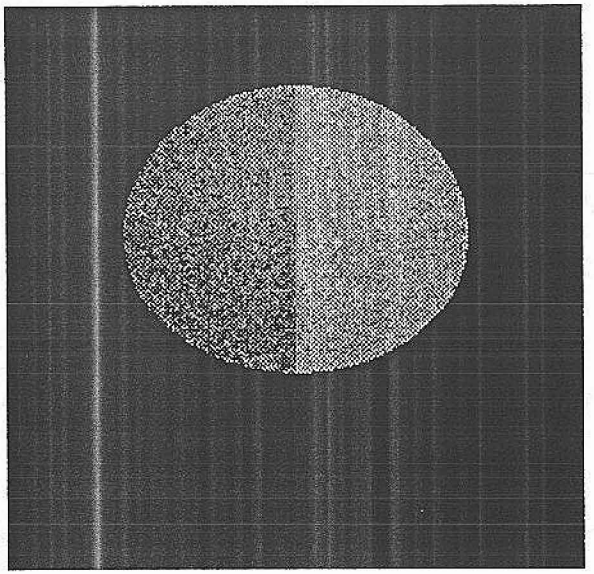


Figure 7.2: A view of the surface reconstructed using the algebraic, class II weakly coupled fusion algorithm.

has already been performed, and that we have a module for computing the sign of γ. We took, for the experiment, $m = 15$, $\hat{s} = (-1, 0, 1)/\sqrt{2}$, $\hat{k} = (0, 0, 1)$. It was assumed that the noise in the measurements of the specular and Lambertian image components was white zero mean Gaussian with variances $\sigma_s^2 = .05$ and $\sigma_l^2 = .025$. Note that the Lambertian measurement is less noisy than the specular measurement. The algorithm, however, does not use this information in any way, as both measurements are absolutely required and we cannot trade one off for the other.

The original shaded image (containing both noisy specular and Lambertian components) is shown in figure 7.1. The resulting surface obtained from the algorithm is shown in figure 7.2. Note that the surface has been "reconstructed" only over a fraction of the surface visible in the image. This is due to the fact that the specular

component E_s becomes shadowed before the Lambertian component does. Since both image components are required for the surface reconstruction process, the surface normals are not computed in the specular shadow area, even though the surface is visible due to the Lambertian component. The view of the reconstructed surface is computed for a light source direction, $\hat{s}_2 = (-1, 1, 1)/\sqrt{3}$, different than the one used to get the data, so that any errors in the surface normals would be apparent. The reconstruction is very poor and is essentially random. The step in brightness is due to the *a priori* imposition of the sign of γ (i.e. γ was fixed to be positive for $y > 0$ and negative for $y < 0$). Since the magnitude of γ is essentially random, the spatially correlated fixing of the sign of γ gives rise to the observed difference in the mean of $\hat{n} \cdot \hat{s}$ for positive and negative y.

The magnitude of the error in the surface normal (i.e. $|\hat{n}(\vec{x}) - \hat{n}_{actual}(\vec{x})|$) over the reconstructed surface is significant and is shown in figure 7.3. White areas indicate high error magnitude (maximum error was a very high 1.22) while dark areas correspond to regions of low error magnitude (minimum error was 0.0). Observe in the figure how the error in the surface normal grows as one moves away from the specular reflection (where E_s decreases) and as one moves towards the limb of the sphere (where E_l decreases).

One problem that was encountered in the application of the algebraic algorithm to shape recovery from noisy images was a lack of stability. This lack of stability is due to the fact that the range of the mapping R from surface normals $\hat{n}$ to image components E_s, E_l does not cover the area $(0, 1) \times (0, 1)$. Thus, there are pairs of E_s, E_l values in the area $(0, 1) \times (0, 1)$ for which there is no corresponding surface normal. For such (E_s, E_l) pairs the value of γ will be imaginary. Figure 7.4 shows (in black) the region in the feasible $E_s \times E_l$ space for which γ is real and corresponding surface normal vectors exist.

In a noise free situation the pathological image component pairs that do not have corresponding surface normals will not be observed. If the values of E_s and E_l are noisy, however, then these values can

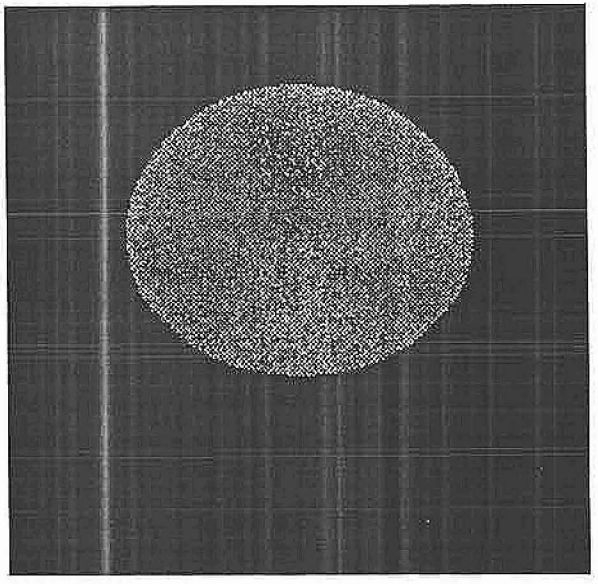
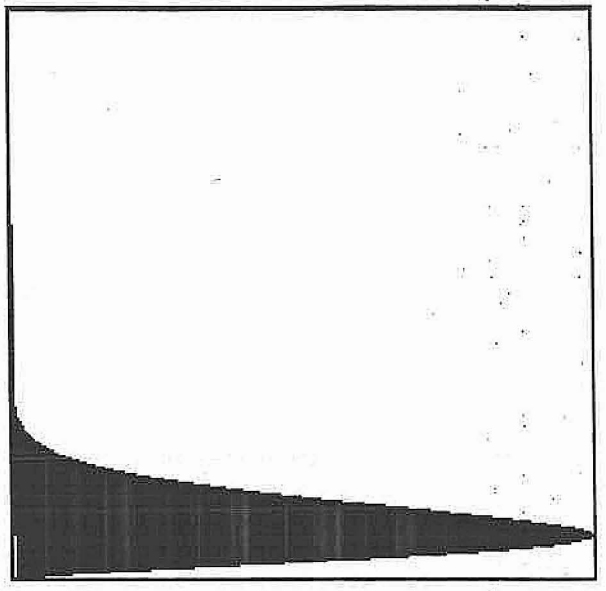


Figure 7.3: The error in the magnitude in the surface normal obtained using the algebraic fusion method.

be observed. It is evident that these pathological values are most likely to arise when the surface normals map to E_s, E_l values near the boundary of the region containing allowable values.

In order to obtain the reconstruction shown in figure 7.2, where surface normal vectors were found for all image points, even those with pathological E_s, E_l pairs, we used the heuristic of finding the E_s, E_l pair, closest to the measured E_s, E_l pair, that lay within the allowable region (i.e. the black area in figure 7.4). This E_s, E_l pair was then used to compute the surface normal vector. This heuristic seems to provide acceptable results, as indicated by the quality of the reconstruction shown in figure 7.2, but is troubling when we consider the stability of the method. It is evident that, for E_s, E_l pairs near the boundary of the allowable region in figure 7.4, a small perturbation of the E_s or E_l values can result in a computed surface



Vertical Axis = Lambertian Component (top = 0, bottom = 1)
Horizontal Axis = Specular Component (left = 0, right = 1)

Figure 7.4: The range of the mapping from surface normals to specular and Lambertian reflectances; black areas indicate surface normals for which there is a corresponding pair of reflectance components.

normal that is quite a bit different than the actual surface normal. Thus the algorithm is ill-posed in the sense that the solution does not always vary smoothly with the input.

The shape from shading algorithm we have just described illustrates a major drawback of all algebraic fusional algorithms. In such algorithms *all* sources of information are absolutely required. There is no way to reduce the dependence of the solution on a unreliable source of data. Thus, in situations where one, or more, of the sources of data is very unreliable, the output of the module will likewise be very unreliable. In order to be able to minimize the effect of an unreliable source of data in such a case we must replace the unreliable

source of data with some other, more reliable, source of information. Since, in the application that we are examining, the specular and Lambertian image components are the only scene dependent sources of information we have available, any additional sources of information must be in the form of *a priori* (scene independent) constraints. Embedding these *a priori* constraints leads us to a class III weakly coupled fusion implementation of the shape from shading algorithm.

7.2 A CLASS III WEAKLY COUPLED FUSION IMPLEMENTATION

As we have noted, there is no way, in general, to embed more constraints in an algebraic formulation of a fusional problem than are needed to obtain a unique solution. A Bayesian or energy function formulation of data fusion, however, has the advantage over the algebraic method that we can add additional constraints on the solution. This approach will result in an algorithm which is more robust away from specularities than the algebraic algorithm that was described in the preceding section.

To show how we can implement such an algorithm let us begin by casting the algebraic solution of the previous section as an energy function minimization problem. Consider:

$$\hat{n} \leftarrow \min_{\hat{n}} \int \left[(E_L - \hat{n} \cdot \hat{s})^2 + (E_s^{\frac{1}{m}} - \hat{k} \cdot \hat{h})^2 + \mu(||\hat{n}|| - 1)^2 \right] dA \tag{7.18}$$

This equation is similar to the one described by Horn and Brooks [64]. The third term in the integrand is required to maintain the length of the surface normals near unity.[1]

[1] Alternatively, one can use a Lagrange multiplier term

$$\mu(\hat{x})(||\hat{n}|| - 1)$$

to impose the unit normal constraint exactly. The Lagrange multiplier $\mu(\hat{x})$ must be computed in some fashion which depends on how the minimization is being performed (for an example see [23]).

It is evident that the global minimum of this energy functional occurs when $\hat{n}(\hat{x})$ is given by equation (5). Hence the solution obtained through the energy minimization process is the same as for the algebraic process. The energy function approach, however, has the advantage that it is possible to embed additional constraints on the problem, which can be used to reduce dependence on unreliable data. For example we could include a smoothness term, of the form $||\nabla\hat{n}||^2 = (|\partial\hat{n}/\partial x|^2 + |\partial\hat{n}/\partial y|^2)$ into the energy functional given above. Alternatively, we could use the "integrability" constraint suggested by Horn and Brooks [64]. This constraint is of the form $(\frac{\partial p}{\partial y} - \frac{\partial q}{\partial x})^2$, where p and q are the x and y slopes of the surface and are related to the unit surface normal by:

$$\hat{n}(x,y) = \frac{(-p,-q,1)}{\sqrt{1+p^2+q^2}} \tag{7.19}$$

and

$$p = -\frac{n_X}{n_Z},\ q = -\frac{n_Y}{n_Z} \tag{7.20}$$

where $\hat{n}(x,y) = (n_X, n_Y, n_Z)$. The integrability term imposes smoothness since only surfaces that are at least C^2 (i.e. twice differentiable) will satisfy this constraint. Integrability can also be obtained from non-integrable surfaces by projecting the non-integrable surface normal estimates onto the nearest integrable surface at each stage of an iterative solution to the minimization process. Such an algorithm was proposed by Frankot and Chellappa [44]. The smoothness terms, however they are implemented, will reduce the effect of noisy inputs on the resulting surface normal values. In addition, since the smoothness constraint allows the solution of the problem even if only one of the sources of data are present (i.e. either E_s only or E_l only), we can *weight* the relative contributions of the input sources of data. Thus, if we know that one source of data is much more reliable than the other, then we can weight that source more highly than the unreliable source. For example, we want to find $\hat{n}(x,y)$ which minimizes:

$$\int [w_l(E_l - \hat{n}\cdot\hat{s})^2 + w_s(E_s - (\hat{k}\cdot\hat{h})^m)^2 + \mu(||\hat{n}|| - 1)^2 + \lambda(||\nabla\hat{n}||^2)]dA \tag{7.21}$$

where w_l and w_s are weights related to the reliability associated with the input data sources E_l and E_s. The weight λ determines the amount of smoothing done on the resulting surface normal field.

The addition of the smoothness term also implicitly solves the problem of the ambiguity in $\hat{n}$ caused by the symmetry of E_s and E_l about the vector $\hat{s}\times\hat{k}$. If we have some boundary conditions then the surface normal vectors with the correct sign of the component in the $\hat{s}\times\hat{k}$ direction will be those that maximize smoothness as well. Notice also that we do not have problem with stability that was present in the algebraic method due to the inversion of the mapping from $\hat{n}$ to (E_s, E_l). In the energy function method an explicit inversion does not need to be performed.

The above energy function formulation can be converted into a Bayesian one through the application of the Gibb's distribution to get:

$$P(\hat{n}|E_l, E_s) = e^{-w_l(E_l-\hat{n}\cdot\hat{s})^2} e^{-w_s(E_s-(\hat{k}\cdot\hat{h})^m)^2} e^{-\mu(||\vec{n}||-1)^2} e^{-\lambda||\nabla\hat{n}||^2} \tag{7.22}$$

The first two terms in the Bayesian formulation can be thought of as the image formation model, while the third and fourth terms are *a priori* constraints which says that surface normal vectors are most likely to have unit length, and that smooth surface normal fields are more likely than irregular fields.

In general, the above energy functional may have many minima and it may be difficult in practice to find the global minimum. For this reason, the algebraic approach, which produces the solution corresponding to the global maximum, may be preferable. It is not always possible, however, to solve a fusional problem analytically. In addition, the sensitivity of algebraic approaches to noise in the input data may be unacceptably high. In such cases, the Bayesian or energy function minimization approaches must be used.

The energy functional based fusion algorithm described above was run on the same synthetic sphere data as was the algebraic algorithm in the previous section. The iterative approach described in [23] is used to minimize the energy functional as follows. First, the Euler-Lagrange equations for the surface normal components associated with the energy functional are found to be:

$$w_s(E_s - R_s(\hat{n}))\frac{\partial R_s}{\partial \hat{n}} + w_l(E_l - R_l(\hat{n}))\frac{\partial R_l}{\partial \hat{n}} + \lambda\nabla^2\hat{n} - \mu\hat{n} = 0 \tag{7.23}$$

where

$$\frac{\partial R_l}{\partial \hat{n}} = \hat{s} \tag{7.24}$$

and

$$\frac{\partial R_s}{\partial \hat{n}} = [2\hat{k}(\hat{n}\cdot\hat{s}) + 2\hat{s}(\hat{n}\cdot\hat{k})]m[2(\hat{n}\cdot\hat{k})(\hat{n}\cdot\hat{s}) - (\hat{s}\cdot\hat{k})]^{m-1} \tag{7.25}$$

We then discretize the Euler-Lagrange equations and rearrange according to the procedure given in [23] to yield the iteration equations:

$$\vec{m}_{ij}^{(k+1)} = \bar{n}_{ij}^{(k)} + \frac{\epsilon^2}{4\lambda}[w_s(E_{s_{ij}} - R_s(\hat{n}_{ij}^{(k)}))\frac{\partial R_s(\hat{n}_{ij}^{(k)})}{\partial \hat{n}} + w_l(E_{l_{ij}} - R_l(\hat{n}_{ij}^{(k)}))\frac{\partial R_l}{\partial \hat{n}}] \tag{7.26}$$

$$\bar{n}_{ij}^{(k)} = \frac{1}{4}(\hat{n}_{i,j+1}^{(k)} + \hat{n}_{i,j-1}^{(k)} + \hat{n}_{i+1,j}^{(k)} + \hat{n}_{i-1,j}^{(k)}) \tag{7.27}$$

$$\hat{n}_{ij}^{(k+1)} = \frac{\vec{m}_{ij}^{(k+1)}}{||\vec{m}_{ij}^{(k+1)}||} \tag{7.28}$$

The purpose of the third equation above is to ensure that the surface normal vector obtained has unit magnitude. Thus, the term involving μ does not appear in the iterative formulation. The indices (ij) index into the spatial lattice of the input arrays E_s and E_l. The index k is an iteration counter.

In this iterative approach it is necessary to have boundary conditions since $\frac{\partial R_s}{\partial n}$ and $\frac{\partial R_l}{\partial n}$ only span a plane. These are typically obtained from the occluding contour (where it is assumed that $n_z = 0$ and

n_x, n_y are given by the x and y components of the perpendicular to the projection of the occluding contour in the image plane) and the self-shadow line. If the reflectance function at the self-shadow line is Lambertian only, then we have that $\vec{n} \cdot s = 0$. This constraint, along with the constraint imposed by the perpendicular to the self-shadow line in the image plane is enough to specify the surface normal along the self-shadow line (this assumes that the image plane is not parallel to the illumination direction, and that the self-shadow boundary lies in a plane that is perpendicular to the illumination direction [72]). We will ignore, for the purposes of this chapter, the problem of distinguishing self-shadow lines from occluding contours.

An important aspect to consider in the implementation of the iterative solution to the Euler-Lagrange equation is the determination of the weights, w_s and w_l. Ideally, we would like w_s and w_l to reflect our confidence in the specular and Lambertian data sources. In practice, however, the value of these weights relative to the λ term affect both the amount of smoothness applied to the reconstructed surface normal field and the stability of the iterative process. As the value of λ decreases the increment added to $\bar{n}_{ij}^{(k)}$ in the iteration equation becomes larger. If λ is too small the iteration will become unstable and convergence will not be attained. If, on the other hand, the value of λ is too large, the increment added to $\bar{n}_{ij}^{(k)}$ will be negligible and the resulting solution will depend only on the boundary conditions. In our example, this would result in a surface that is more like a potato chip than a hemisphere. Thus, we would like to adjust λ so that it is small enough to prevent excessive smoothing of the surface, but still large enough to ensure that the iteration process converges. If we had only one data source, the determination of a suitable value for λ would be fairly straightforward; a trial and error process could be used, at the very least, to determine, for a given weight on the data term, the minimum value of λ for which the iteration converges. With two or more sources of data, however, the problem of determining λ is complicated by the presence of multiple terms being added to $\bar{n}_{ij}^{(k)}$. We must ensure that each of these terms always be less than a certain level, else instability will arise. The simplest way to do this is to set the weights on each of the terms (i.e. w_s and w_l) to be equal

to the inverse of the maximum possible value of the data dependent update term, and then determine λ correspondingly (i.e. by trial and error given these w_s and w_l values). That is, in our case we would have:

$$w_s = \frac{1}{\max|(E_s - R_s)\frac{\partial R_s}{\partial \hat{n}}|} \tag{7.29}$$

and

$$w_l = \frac{1}{\max|(E_l - R_l)\frac{\partial R_l}{\partial \hat{n}}|} \tag{7.30}$$

The max operator in the above equations is taken over all possible normal vectors (i.e all unit vectors) and over the three components of the argument of the max operator (i.e. we use the component with largest absolute value). We can determine these weights by assuming that $\max(E_s - R_s) = \max(E_l - R_l) = 1$ and differentiating $\partial R_s/\partial \hat{n}$ and $\partial R_l/\partial \hat{n}$ with respect to $\hat{n}$ and set the result equal to zero to obtain the maximum value attained by $\partial R_s/\partial \hat{n}$ and $\partial R_l/\partial \hat{n}$. Doing so for the illumination parameters of the example we have been using of the sphere yields:

$$w_l = \sqrt{2} = 1.414 \tag{7.31}$$

and

$$w_s = \frac{[\sqrt{2}]^{14}}{15\sqrt{2}} = 6.034 \tag{7.32}$$

This approach, however, does not allow us the liberty of trading off the Lambertian data over the specular data if it is known that one data source is more reliable than the other. We can, however, use the weights given in the above equations as upper bounds, and use the information regarding the relative reliabilities of the data sources to reduce one of them. For example, if the uncertainty in E_s was twice that of E_l we would want the relative weighting of E_s to be half that of E_l. In order to keep the amount of smoothing as low as possible we would set w_l to be the upper bound give above in equation (7.29) and set w_s to be half of the upper bound given in equation (7.30). Hence for our example in which the variance in the noise added to E_s is 0.05, and the variance of the noise added to E_l is 0.025 we would set $w_s = 3.017$ and keep w_l at its upper bound, 1.414. With

these weights it was found that λ needed to be at least 35 to ensure stability of the iterative process.

Observe that, because we chose λ to be the minimum possible subject to the stability of the iterative process, we have given up the ability to adjust the level of smoothing of the solution as a function of the variance of the noise in the data. That is, in our approach, the same value of λ would be used when there is a lot of noise as when there is little or no noise. This occurs because we are using the smoothing term not to smooth the data, but to regularize, or permit an unique solution for, the shape from shading problem. Thus, even when there is no noise we must still have the smoothing term present. We could raise the value of λ from the lower limit required for stability when the noise level is relatively high. This would, in effect, be placing more faith in the *a priori* constraint, and the class of surfaces that it implies than on the data.

A shaded view of the resulting reconstructed surface is shown in figure 7.5. Compare this surface to that produced by the algebraic algorithm. Note that we have surface normal values over the entire illuminated portion of the sphere, as the smoothness constraint along with the Lambertian component is sufficient to solve for the surface normals in the specular shadow region.

The magnitude of the error, coded by brightness as before, in the surface normal vectors is shown in figure 7.6. The maximum surface normal error magnitude (whitest pixels), after 1100 iterations, is about 0.7, corresponding to an angular error of about 36 degrees. The average error is about 0.075, corresponding to an average angular error of about 5 degrees. Observe that the largest part of the error in the resulting surface surface normals occurs near (but a little bit away from) the occluding edges of the sphere and near the specular shadow boundary, where the specular and Lambertian data are most noisy. The best reconstruction of the surface actually occurs near the maximum of the specular image. The specularity in the view of the reconstruction shown in figure 7.5 is quite distorted, indicating a distortion of the surface. This distortion is large as the specularity,

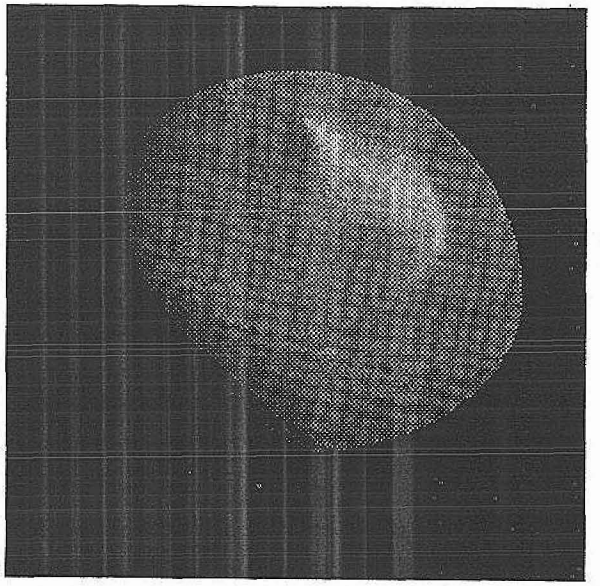


Figure 7.5: A view of the surface reconstructed with the class III weakly coupled fusion method.

for the direction of the light source used to create this view, lies near the specular shadow boundary (for the light source direction used to obtain the data). In the specular shadow region we set $w_s = 0$ and just used the Lambertian data. This explains the sudden drop in the error at the specular shadow boundary.

It might be concluded from this that the error in the recovered surface normals are due to the noise in the specular and Lambertian data, but this is not entirely true. In fact the major portion of the error results from the action of the smoothness constraint that was used. The role of the smoothness constraint is to make "more likely", or give a "lower energy" to, surfaces that have low spatial variation in the components of their normal vectors. There is a class of surfaces that are "most likely", or have minimum energy. We call these surfaces minimal surfaces. The energy minimization process

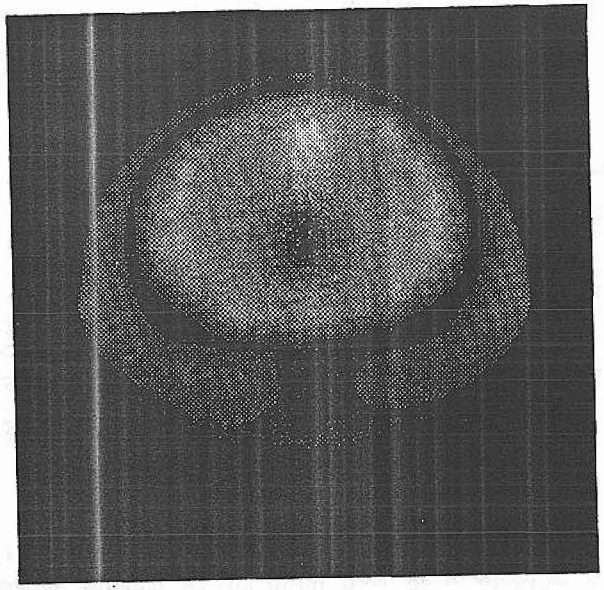


Figure 7.6: The magnitude of the error in the surface normals obtained with the class III weakly coupled fusion method.

pushes the solution surface towards these minimal surfaces, subject to a "repulsive" action due to the image based energy terms. In general, the actual shape of an object may be quite far from that of a minimal surfaces, even though it may be quite smooth. In this case the energy minimization based shape from shading algorithm may produce a solution that has significant error. This behavior is seen in the above example. The errors in the computed shape are mainly due to the deformation of the surface towards a minimal surface.

What is the form of the minimal surface normal field that the smoothness constraint is forcing the solution towards? To determine this we look at the case where $\lambda = \infty$ (i.e. there is no dependence on the data).

In this situation the Euler equations reduce to:

$$\begin{aligned} \nabla^2 n_x &= 0 \\ \nabla^2 n_y &= 0 \\ \nabla^2 n_z &= 0 \end{aligned} \tag{7.33}$$

Both n_x and n_y are independent but n_z is constrained to be $n_z = \sqrt{(1 - n_x^2 - n_y^2)}$. In order for $\nabla^2 n_z = 0$ we must therefore satisfy the following conditions on n_x and n_y (obtained after a large amount of algebraic manipulation):

$$(1 - n_y^2 + 3n_x^2)|\nabla n_x|^2 + 8n_x n_y \nabla n_x \cdot \nabla n_y + (1 - n_x^2 + 3n_y^2)|\nabla n_y|^2 = 0 \tag{7.34}$$

This equation defines a class of minimal surface normal vector fields. This class includes the constant vector field, corresponding to a flat surface, but does not include the vector field corresponding to a sphere (as can be seen by substitution of $\hat{n} = (x/r, y/r, (1-x^2-y^2)/r)$ into the equation above). Therefore, even with noise free data, for non-zero λ the minimum of the energy functional will not result in a spherical surface. In fact, if we run the iterative energy minimization with $\lambda = \infty$, thereby removing the effect of the image data, the resulting surface (which is a minimal surface) results in an error map (deviation from a sphere) that is quite similar in its spatial pattern (at least in the specular shadow region) to that shown in figure 7.6.

As mentioned above, the effect of the smoothness constraint on the solution is to force the solution towards an certain class of surfaces. In general we would want the smoothness constraint to be such that this class of surfaces was as general as possible, and hopefully including the true solution. We should make our constraints as weak as possible while ensuring that we can obtain a unique solution. In this regard, the integrability constraint is a good one as its corresponding class of minimal surfaces is the class of C^2 surfaces, which is very large (and includes the hemisphere without boundary used in our experiments). Alternatively we could adapt our constraint using information about the shape obtained from independent sensory processing modules. These can be used to alter the constraints to achieve consistency in solution between modules. This (strongly

coupled) form of data fusion is discussed briefly in Chapter 8, in the context of active data fusion.

Whatever the drawbacks of the regularization type approach to the solution of the shape extraction problem, it is clearly better than the algebraic approach. The improvements over the algebraic method are mainly due to the effect of smoothing on the surface normal field, the elimination of the stability problems inherent in the algebraic method and in a very small amount to the relative weighting of the Lambertian data over the specular data. In addition the surface over which the surface normal was computed is greater than in the algebraic method since we can obtain solutions even where there is no specular information, which was not possible in the algebraic method.

In order to improve our surface reconstruction even further we need to use a strongly coupled fusion algorithm that will control the weighting applied to the input data as a function of its local reliability, rather than only globally applying weighting values.

7.3 A STRONGLY COUPLED APPROACH TO POLYCHROMATIC SHAPE FROM SHADING

The weakly coupled fusion algorithm described in the previous section uses only the information about the reliability of the input data E_s and E_l in determining weights for the two sources of data. It did not use any information regarding the *sensitivity* of the shape from shading algorithm to errors in the inputs. It therefore can have problems in areas of the image where the sensitivity of the shape from shading algorithm to errors in the input data was high. Ideally we would like to reduce the weighting of the data consistency terms when the sensitivity to errors in the data is high, as well as when the errors in the data are large. We could consider having a separate module

whose task was to take in the values of E_s and E_l and compute the shape from shading algorithm sensitivities. These sensitivities would then be fed into the fusional module. The sensitivities would be used by the fusional module to *adaptively* determine weights to be applied to the data consistency terms for E_s and E_l.

It is evident that, in principle, such a fusional algorithm would be considered to be strongly coupled since the data from the sensitivity module is strongly coupled to the shape from shading module. Practically speaking, however, the Lambertian data and the specular data are still weakly coupled as the fusion does not affect the extraction of either the Lambertian or specular image components (one could, however, think of an algorithm by which this did occur). The strong coupling occurs between the sensitivity module and the shape recovery module (and not between the sensitivity module and the Lambertian or specular extraction modules). This is an important distinction to be made.

The energy function formulation of the strongly coupled approach to polychromatic shape from shading is identical to that of the previous section (equation (7.21)), however, now the weights w_s and w_l are longer constant over the image but are now some function of the sensitivities to noise that were derived earlier (i.e. equations 7.15 and 7.16). We must still be careful to ensure that the iterative solution to the Euler-Lagrange equation is stable, by proper scaling of the weight values. As in the previous section we do this by setting the maximum value of w_s and w_l to the values given in equations (7.29) and (7.30) and reduce these weights by an amount related to the relative uncertainty in the data. It is not clear exactly what form should be used to express the weight values in terms of the sensitivities, save for a general rule that the weights should decrease as the noise sensitivities increase. We use the following law:

$$w_s(x,y) = w_{s_{max}} \left[\frac{\min(\sigma_l^2, \sigma_s^2)}{\sigma_s^2} \right] \frac{1}{1 + \log\left(1 + \left|\frac{\partial \hat{n}(x,y)}{\partial E_s}\right|\right)} \qquad (7.35)$$

$$w_l(x,y) = w_{l_{max}} \left[\frac{\min(\sigma_l^2, \sigma_s^2)}{\sigma_l^2} \right] \frac{1}{1 + \log\left(1 + |\frac{\partial \hat{n}(x,y)}{\partial E_l}|\right)} \quad (7.36)$$

where $\partial\hat{n}/\partial E_s$ and $\partial\hat{n}/\partial E_l$ are given by equations (7.15) and (7.16), and where $w_{s_{max}}$ and $W_{l_{max}}$ are the upper bounds given in equations (7.31) and (7.32). The log function was used to compress the sensitivity functions in order to retain some data dependent terms in the energy function near the boundaries, otherwise the surface would be excessively smoothed near the boundary of the sphere and near the shadow line.

To see more clearly the strong coupling that occurs in this method consider the Bayesian formulation of the fusion process. Specifically, we can write:

$$p(\hat{n}|E_s, E_l) = \frac{p(E_s, E_l|\hat{n})p(\hat{n})}{p(E_s, E_l)} \quad (7.37)$$

where we have the following expression for $p(E_s, E_l|\hat{n})$:

$$p(E_s, E_l|\hat{n}) = e^{-[(E_s^{\frac{1}{m}}(x,y) - R_s(\hat{n}))^2 w_s(x,y) + (E_l(x,y) - R_l(\hat{n}))^2 w_l(x,y)]} \quad (7.38)$$

The *a priori* density is assumed to reflect surface smoothness and unit surface normal magnitude constraints as follows:

$$p(\hat{n}) = e^{-[\mu(||\hat{n}||-1)^2 + \lambda||\nabla\hat{n}||^2]} \quad (7.39)$$

It is seen that the image formation model assumed at a given position (x, y) depend on the value of the weights $w_s(x, y)$ and $w_l(x, y)$. These weight values come from the module which computes the shape from shading sensitivity, so it is clear that the sensitivity module is strongly coupled to the shape from shading module as the outputs of the sensitivity module alters the image formation model used in the fusional process, *in a data dependent manner.*

To compare the performance of the adaptive fusion algorithm with the two weakly coupled algorithms described earlier, we ran the adaptive algorithm on the sphere data that was used in the previous sections. The exact iterative equations are used as in the implementation of the energy functional algorithm described in the previous section. In the present case, however, the weights w_s and w_l were functions of position as indicated by equations (7.35) and (7.36). The view of the reconstructed surface is shown in figure 7.7, and the surface normal error magnitude map is shown in figure 7.8. We find a significant improvement, as compared with the weakly coupled approaches, towards the boundary of the specular zone due to the reduction of w_s in this region. The maximum surface error normal error magnitude for this algorithm was 0.17 after 1500 iterations, corresponding to a maximum angular error of 9.6 degrees. The average surface normal error magnitude was 0.057, corresponding to an average angular error of 3 degrees. The effect of the specular data is felt most strongly near the specularity. To see this, we compare the surface normal error magnitude obtained with the adaptive fusion algorithm with the errors produced by running the adaptive algorithm

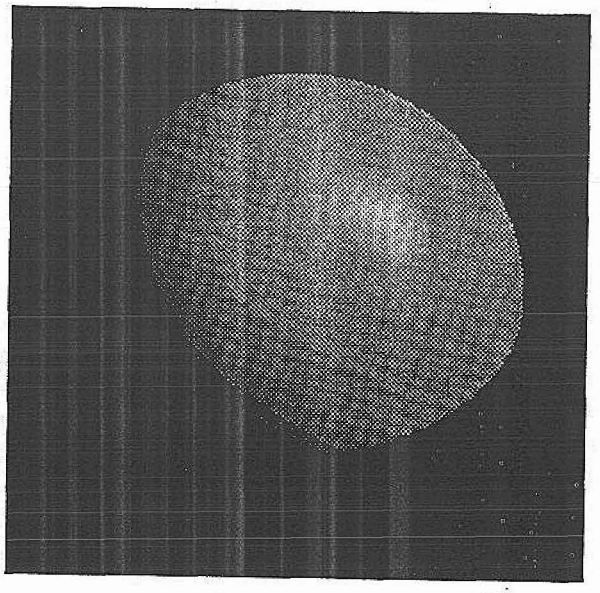


Figure 7.7: A view of the surface reconstructed with the strongly coupled fusion method.

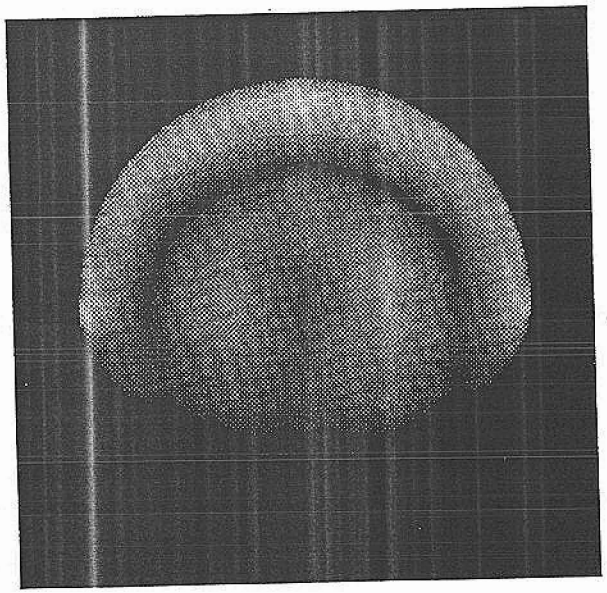


Figure 7.8: The magnitude of the error in the surface normals obtained with the strongly coupled fusion method.

with w_s fixed to be zero everywhere (i.e. no influence by the specular component, only the Lambertian component). The error was slightly lower near the specularity when we take the specular data into account than when we do not. The maximum surface normal error magnitude in the Lambertian only case, after 1500 iterations was 0.17, while the average surface normal error magnitude was 0.062.

7.4 FUSION OF IMAGE FORMATION MODELS

The shape from shading algorithm described in the previous section involved the adaption of an image formation model based on data from an external sensory module (in this case the sensitivity

module). In this section we will look at the shape from shading problem from a slightly different viewpoint and come up with a different algorithm for solving it.

In the approaches described in the previous sections we had, as data, the Lambertian and specular components of the light reflecting off of the object's surface. We then had a single image formation model which described how these two components are produced from a given surface normal vector. In the approach taken in this section we assume that we have not performed the decomposition of the reflected light into its Lambertian and specular components, but rather have only the intensity of the reflected light as our data. In order to handle the presence of both specular and Lambertian components in the data we will specify two possible image formation models, one that describes the specular reflection and the other describing the Lambertian reflection. The task of the fusion process, then, is to decide which model is appropriate.

One can see that this approach is related to the idea of multi-subjective probabilities that was introduced in Chapter 2. We can think of two observers viewing the same data (the intensity of the light reflecting off of the object) and having different subjective ideas of how the image formation process operates. One will model the surface as purely Lambertian, while the other models the surface as purely specular. In order to arrive at a solution we must specify a single, possibly fused, image formation model. Thus we must either believe one observer or the other, or come to some sort of consensus.

For the application of shape from shading it is a fairly good assumption that, in the neighborhood of a specularity, the intensity of the specular component exceeds that of the Lambertian component, and away from the specularity the Lambertian component dominates. Thus, in doing our fusion of the image formation models, it suffices to select one or the other of the models and do not need to attempt to reach some consensus (e.g. half-Lambertian, half-specular). Note that as a byproduct of this shape from shading process we obtain a *segmentation* of the image into regions which are primarily

Lambertian and regions which are primarily specular. Observe that the segmentation process and the shape from shading process are therefore strongly coupled since the solution of the shape depends on the solution of the segmentation.

The image-formation model fusion approach can be implemented in an energy function minimization framework as follows. Consider the energy function shown below:

$$\int [(E - \hat{n} \cdot \hat{s})^2 w_l + (E - (\hat{k} \cdot h)^m)^2 (1 - w_l) + \mu(||\hat{n}|| - 1)^2 + \lambda(||\nabla \hat{n}||^2)] dA + \int_\Gamma C dl$$

The factor w_l is an "area" process (as compared to a "line" process) which is one in regions of the image whose brightnesses are considered to be mainly due to Lambertian reflectance, and is zero elsewhere (i.e. in the specular regions). Note that, unlike the line processes described earlier, the area processes are indicating places where the image formation model is invalid or is changing, rather than places where the *a priori* constraints (e.g. smoothness) break down. The third and fourth terms in the energy function are the smoothness and unit magnitude constraints on the surface normal field as seen earlier. The smoothness constraint is required for regularization purposes. The last term is intended to prevent many small, or isolated, specular or Lambertian regions from being formed. This term is proportional to the total length of the boundary elements, Γ of $w_l(\vec{x})$. This causes the solution process to prefer solutions that contain large collections of specular or Lambertian points.

The solution process involves jointly minimizing the above energy function with respect to the surface normal field $\hat{n}(\vec{x})$ and the area process $w_l(\vec{x})$. One can rewrite the energy function in a way which makes the strongly coupled nature of the image formation model more apparent:

$$\int (E - w_l(\hat{n} \cdot \hat{s}) - (1 - w_l)(\hat{k} \cdot \hat{h})^m)^2 dA + \int [\mu(||\hat{n}|| - 1)^2 + \lambda(||\nabla \hat{n}||^2)] dA + \int_\Gamma C dl$$

One could explicitly decouple the segmentation process and the shape determination process. Such an algorithm would operate by first performing the segmentation by some means, and then using the knowledge of which regions are specular and which are Lambertian to select a suitable image formation model for the Bayesian determination of the surface normals. We have seen how to do the Bayesian shape determination given a suitable image formation model and *a priori* constraint; we now have to examine ways in which the segmentation can be done. The most straightforward approach in the Bayesian paradigm is to find the segmentation $w_l(\vec{x})$ which minimizes the following energy:

$$E_s = \int (E - w_l(\hat{n} \cdot \hat{s}) - (1 - w_l)(\hat{k} \cdot \hat{h}_l)^m)^2 dA + \int_\Gamma C dl$$

where now both $E(\vec{x})$ and $\hat{n}(\vec{x})$ are known. Observe that the requirement that the surface normal vectors be known means that we cannot get rid of the strongly coupled and recurrent nature of the data fusion. This approach to segmentation is similar to that obtained by using the techniques of statistical communication theory [109]. Here we consider the segmentation problem as one of deciding, in the presence of noise, whether we have one signal or another (i.e. the binary decision problem [109]). As mentioned in chapter 2 the optimal Bayesian decision rule, $\delta(\gamma|\vec{d})$ is that which minimizes the "average risk", where the average risk is defined as [109]

$$R(p(S), \delta) = \int_\Omega dS p(S) \int_\Gamma d\vec{d} p(\vec{d}|S) \int_\Delta d\gamma C(S, \gamma) \delta(\gamma|\vec{d})$$

In the above equation S is the "signal" (i.e. either Lambertian or specular), $p(S)$ is the *a priori* probability of the signal, $\vec{d}$ is, as usual, the (noisy) data, γ is the decision (e.g. $\gamma = 1$ for specular, -1 for Lambertian), $p(\vec{d}|S)$ is the "image formation model" relating how a signal (either specular or Lambertian) gives rise to the data (and hence depends on $\hat{n}(\vec{x})$), and $C(S, \gamma)$ is the cost associated with making the decision γ when the signal is actually S. For the binary case we can replace the integration over the signal space (Ω) by a sum of two terms corresponding to the two elements in the signal space. The *a priori* density on S now is a sum of two delta functions, $p(S) = q$

if S = specular, $p(S) = p$ if S = Lambertian. Similarly, the decision space Δ contains only two possible decisions ($\gamma = +1, \gamma = -1$), and so the integral over the decision space can also be reduced to a sum of two terms. There are four possible combinations of arguments to the cost function and we define $C_{1-\alpha} = C(S$=specular$, \gamma = +1)$, $C_{\alpha} = C(S$=specular$, \gamma = -1)$, $C_{1-\beta} = C(S$=Lambertian$, \gamma = -1)$, and $C_{\beta} = C(S$=Lambertian$, \gamma = +1)$. We now can write the average risk as [109]

$$R(p, q, \delta) = qC_{1-\alpha} + pC_{1-\beta} + q\alpha(C_{\alpha} - C_{1-\alpha}) + p\beta(C_{\beta} - C_{1-\beta})$$

where

$$\alpha = \int_{\Gamma} p(\vec{d}|S = specular)\delta(\gamma = -1|\vec{d})d\vec{d}$$

and

$$\beta = \int_{\Gamma} p(\vec{d}|S = Lambertian)\delta(\gamma = +1|\vec{d})d\vec{d}$$

The α and β are related to the probability of making an error in the decision process.

The optimum Bayes decision rule can be shown to be [109] Decide specular when $\Lambda <$K, and Lambertian otherwise. Λ is the *generalized likelihood ratio,* and is given by

$$\Lambda = \left(\frac{p}{q}\right)\left(\frac{p(\vec{d}|S = Lambertian)}{p(\vec{d}|S = specular)}\right)$$

Note that Λ is always non-negative. K is a decision threshold, and is related to the costs as follows,

$$\mathrm{K} = \frac{C_{\alpha} - C_{1-\alpha}}{C_{\beta} - C_{1-\beta}}$$

We can determine the form of Λ once we have specified the two image formation models, $p(\vec{d}|S = specular)$ and $p(\vec{d}|S = Lambertian)$. The determination of the threshold K requires the specification of the cost of making a mis-classification. Typically we will want to use these costs to embed our constraint concerning the suppression

of small regions in the segmentation. This will require that the costs be dependent on the spatial structure of the solution, and hence implies that an iterative procedure is required to embed this constraint through the specification of the costs. In the absence of this constraint we could set $C_{1-\alpha}$ and $C_{1-\beta}$ to be zero (i.e. no cost assessed for correct segmentation) and set $C_{\alpha} = C_{\beta}$. Then K=1. We can then rephrase the decision process as: decide specular when $\log \Lambda < 0$ and decide Lambertian otherwise. The quantity $\log \Lambda$ is seen to be

$$\log \Lambda = [(E - \hat{n} \cdot \hat{s})^2 - \log p] - [(E - (\hat{h} \cdot \hat{k})^m)^2 - \log q]$$

This corresponds to the difference in the energies between the two possible solutions in the energy function formulation. The terms $-\log p$ and $-\log q$ represent the *a priori* constraint of the energy function formulation (i.e. the *a priori* expectation of a piece of surface being either specular or Lambertian - the minimum area constraint could possibly be embedded here by making p and q be dependent on the surface normal field $\hat{n}(\vec{x})$). We can see the relationship between the energy function formulation and the optimal binary decision making formulation of the segmentation process. If the energy corresponding to the assumption of specular image formation model is less than the energy for the assumption of the Lambertian image formation model then $\log \Lambda$ will be less than zero, and we would decide on the specular solution.

The shape from shading method described here can be extended to the case where there are a multiple number of possible image formation models (each corresponding to a different surface shading law). A more complex decision rule would need to be used in place of the simple threshold, but the general approach would be the same.

The method described in this section is a strongly coupled one, as it uses the segmentation obtained from the image to provide the proper image formation model to use in the shape determination process. It is also a recurrent method, as it requires knowledge of the surface normal field in order to perform the segmentation. It is possible to specify alternative methods which are not recurrent, by requiring that the segmentation be done without knowledge of the

surface normal field. To do so in the Bayesian decision framework is difficult because of the non-parametric form of the problem [109]. That is, the conditional probability $p(\vec{d}|S)$ is not fully determined by specifying S (since $\hat{n}(\vec{x})$ is also needed), which makes the derivation of a suitable decision rule difficult, if not impossible. There are algorithms, however, that are not based on the Bayesian approach, which can do the segmentation using only the intensity data. An example is the method proposed by Brelstaff and Blake [22] which bases the segmentation on measures of consistency of the intensity with Lambertian constraints.

Like all recurrent fusion methods we must concern ourselves with the dynamics of the implementation. Can we come up with an iterative algorithm which will converge, and which will converge to the right answer? One possible approach, which has not been tested or analyzed yet, is to start by assuming the surface is Lambertian everywhere. This is a reasonable assumption for smooth surfaces that have a high coefficient of specularity (m in the above equations). One would then run the shape from shading part of the module using the Lambertian image formation module. The surface normal field so obtained can be used in the Bayesian decision process to modify the segmentation. Based on the new segmentation the shading process would be repeated, giving a new surface normal field, and then a new segmentation, and so forth.

Note that in the specular regions the surface normals computed during the first pass will be incorrect, due to the invalid assumption of a Lambertian image formation model. It is not guaranteed, however, that the segmentation based on these incorrect values will be correct. Thus some points in the specular region will remain misclassified as Lambertian. Some points may be correctly classified. In addition, it is expected that the surface normals in the Lambertian region will be computed correctly, so that on the whole, the segmentation will improve at each iteration and the process will converge to the correct solution. The only possible problems will occur near the border of the specular and Lambertian regions, where the surface normals will be computed incorrectly during the initial phases. The smoothness constraint may propagate these errors into the Lambertian region, causing a degradation of the segmentation. The smoothness constraint might, however, have the opposite effect, propagating the correct surface normals into the specular region, thereby improving the segmentation. Thus the answer to the question of whether the algorithm will converge or not will depend on the precise form of the smoothness constraint.

7.5 CHAPTER SUMMARY

- We presented a series of methods for determining the shape of an object from knowledge of the specular and Lambertian components of the light reflecting off of the object. Our first method was a standard weakly coupled fusional approach that allowed and algebraic solution. It was found the performance of the algebraic approach was poor due to high sensitivity of the solution to errors in the data, as well as due to the ill-posedness resulting from the mapping from surface normals to intensity components not being onto (that is, there are some specular-Lambertian intensity pairs that do not have associated surface normals).

- To get around the problems of the algebraic fusion method, we proposed an energy functional based weakly coupled fusion method. This approach combines two energy terms that measure how consistent the computed surface normals are with the specular and Lambertian components, and a term which imposes a smoothness constraint. The smoothness constraint is needed to stabilize the solution in regions where the noise in the components is high or where one of the components is missing due to shadowing. The energy functional was minimized by applying the calculus of variations, which specifies the Euler equation that the minimizing solution must satisfy. This partial differential equation was discretized and solved with an iterative relaxation scheme. As with all iterated systems, one must be careful to prevent instabilities. This provides constraints on

the values of λ that we can use.

- We introduced a strongly coupled method which used estimates of the sensitivity of the recovered shape to noise in the image components to adaptively weight the specular and Lambertian consistency terms in the energy functional. This approach had the benefit of reducing the dependence of the solution on the data when the data was noisy, or when the solution process was most sensitive to the noise. The fusion was a form of feedforward image formation model adaption strongly coupled fusion, as the data from the sensitivity is used to alter the determination of the shape. The computed shape has no effect on the determination of the sensitivity, and so the fusion is not reciprocal strong coupling.

- The final approach that was described involved fusion of image formation models, rather than the fusion of data. This technique was effectively a fusion of a segmentation module and a shape from shading module. The segmentation was that of partitioning the image into regions of mainly specular reflectance and mainly Lambertian reflectance. Two different methods were proposed. The first is essentially a coupled Markov Random Field approach wherein an energy functional is minimized with respect to the surface normal field and a "area process". The second method decouples, somewhat, the segmentation and surface normal computation processes. A technique based on Bayesian binary decision processes was described for implementing the segmentation. This shape from shading method is seen to use strongly coupled recurrent fusion.

Chapter 8

Temporal Aspects of Data Fusion

Time is an important, and all too often ignored, aspect of the operation of sensory systems. Most of the attention paid to temporal aspects of sensory information processing has been regarding the measurement of temporal properties of the environment, such as the velocity of moving objects, and relatively little attention has been given to the role in sensory information processing of temporal constraints on both the world models and the sensory processing algorithms themselves.

The configuration of physical structures in any useful environment or world will change over time. This time variation makes the difficult problem of inverting the world-image map even more difficult, since the dimensionality of the space of possible world configurations is increased. Thus, in order to effectively estimate the changing parameters of a changing world one must impose more constraints than those that are required in a static world. The obvious way to obtain these additional constraints is through sensory measurements taken at multiple points in time. This raises the question of when these additional measurements should be taken, in order to suita-

bly constrain the solution. To determine this, we must model the temporal properties of the world. Clearly, if the parameters describing the world configuration change relatively slowly over time, then measurements can be taken at relatively large intervals in time.

In addition to determining temporal sampling schemes, knowledge of the temporal model (of the environment) allows one to derive ways in which the sensory information obtained at different times can be combined, or fused. We will discuss a class of algorithms known as *temporal coherence* methods, which assume a specific form of temporal model (i.e. that the world parameters change smoothly).

When one considers the operation of sensory systems in time, one cannot ignore the temporal constraints imposed by the computational system. For example, sensory processing algorithms take time to produce their results, and one must examine the effect of this latency on the reliability of the results. In addition, this computational latency may be different for different processing modules, so that the outputs of these modules need to be registered temporally before they can be fused.

8.1 A TEMPORAL COHERENCE EDGE DETECTOR

One can consider the measurements taken by a particular sensor at different times as coming from different sensors. These measurements can then be fused, as if we were fusing the outputs of independent sensory modules. In performing this fusion one has available a powerful constraint, the temporal coherence constraint. This is a constraint on the image formation model (actually the "system" model described in chapter 2), which states, in its simplest form, that the world parameters to be estimated will not change over time.

More complex temporal coherence constraints allow the system to

change in a less restricted, but still known, manner. The state transition matrix used in the specification of a Kalman filter approach reflects one way of embedding this constraint. In general, the specification of a temporal coherence constraint permits the construction of a recurrent, or recursive, parameter estimation procedure. This procedure need not be a Kalman type filter. Other approaches are possible.

In the remainder of the chapter we present an edge detection method that uses a temporal coherence constraint. The estimation process derived from this constraint is based on binary decision theory, rather than on Kalman filtering. The operation is similar, in that as time progresses we change the prior model based on the measured data (and hence the fusion is strongly coupled).

For the edge detection method presented here we assume that candidates for edges are provided by a suitable edge detection method (of which there are a large, but finite, number to be found in the computer vision literature). Typically the images on which edges are to be detected are noisy, causing noise or false edges to appear in addition to edges which correspond to actual physical structures in the scene. We would like to have some way in which true edges can be distinguished from noise edges. Temporal data fusion, in which information form successive images in time are fused, can be used to do this. Initially an edge is assigned a small probability of being a true edge. However, if the edge persists over time then its chances of being due to noise are reduced. Moreover, if the edge also undergoes smooth motion during that period the measure of reliability is increased. Conversely, if an edge moves in a random fashion, or disappears, then the confidence in the edge being a true edge is reduced.

One can put this edge detection method in a rigorous framework by noting that one is looking for statistical evidence that a given edge candidate is a true edge or due to noise. Every bit of information that we obtain will reduce the confidence interval of the hypothesis that a candidate is a true edge. For example, we can track the

change in position, Δx, of a candidate edge over time. The mean of the Δx random variable will be proportional to the velocity of the edge, and its variance will be inversely proportional to the signal to noise ratio. Thus one could apply a statistical hypothesis test to the estimates of the mean and variance of the increments in the position over time of a candidate edge. As time goes on the confidence interval of the decision will be reduced as the accuracy of the mean and variance increases. For weak edges (relative to the noise level) one may have to integrate over many time samples before the confidence in the decision as to whether the candidate is an edge or not is high enough. For strong edges this integration time may be quite short. In addition, if the edge is moving, one can use the additional information provided by the estimation of the mean to make the decision as to the authenticity of the edge. This means that, for a given signal to noise ratio, a moving edge will be distinguished from noise more rapidly than a stationary edge. Furthermore the faster the edge moves, the easier it is to distinguish the difference in the mean of the increment Δx from zero. Hence a fast moving edge will be detected more quickly than a slow moving edge.

We can perform a similar process to that outlined above in a Bayesian data fusion approach. We consider the measurements of the edge candidate position increments Δx at the different time slices t_i to be provided by independent (but identical) measurement systems. Thus, we will be fusing the data from N different data sources, where N is the number of time frames examined.

The process is essentially that of a class I weakly coupled algorithm where the information about the increments Δx taken at different time frames are assumed to be independent and can therefore be combined. The algorithm is formulated in a Bayesian framework as follows. We take as our space of possible solutions the space containing the two elements, true-edge and noise-edge. We can then determine the conditional probabilities of a true-edge given the data, and of a noise-edge given the data. Denote the true-edge solution as f_1, the noise-edge solution as f_2, and the increment Δx at time step

i as d_i. Then we can write:

$$P(f_1|d_1, d_2, ..., d_N) = \prod_{i=1}^{N} \frac{P(d_i|f_1)P(f_1)}{P(d_i)}$$

and

$$P(f_2|d_1, d_2, ..., d_N) = \prod_{i=1}^{N} \frac{P(d_i|f_2)P(f_2)}{P(d_i)}$$

The prior distributions on $P(f)$ just measure the likelihood of a given edge being due to noise or being a true edge. This will be some function of the density of edges in the signal and the noise bandwidth[1]. The $P(d_i)$'s merely scale each of the conditional probabilities and so they play no role in determining which conditional probability has a higher value.

8.1.1 DETERMINATION OF THE CONDITIONAL DENSITIES

The conditional densities $P(d_i|f_1)$ and $P(d_i|f_2)$ are the important components of the Bayesian formulation of the problem. In general the determination of these densities is very difficult and in most cases only approximations can be obtained. For a review of some of the approaches that have been taken to this and similar problems see [17, 91]. However, for particular choices of the defining parameters, these densities can be measured experimentally and used in lookup tables, or determined by approximate analytical formulae.

In general, however, lookup tables will not be sufficient. This is because we will not know some of the parameters, such as step edge amplitude or step edge velocity. It would be impractical to construct lookup tables for each possible set of parameter values. Therefore we would like to try and come up with analytical approximations

[1] If the noise bandwidth is high then the density of edges (zero crossings) due to noise will be high [131]

to the conditional probabilities which we can use to estimate the reliability of our temporal coherence edge detector. We will examine the one dimensional case for simplicity sake and make the following assumptions:

- We have a single step edge with amplitude A.
- White noise with variance σ_n^2 is added to the step edge.
- The resulting noisy signal is then differentiated twice and filtered with a Gaussian lowpass filter with a space constant σ_f.
- We associate edges with zero crossings of the filtered noisy signal.
- We track edges by matching an edge at t_1 with the edge at t_2, having the same sign (of the edge slope), that is closest to the edge at t_1.

We will proceed by deriving an approximation to $p(d|f_1)$ (i.e. for the step edge plus noise) and will then obtain an approximation to $p(d|f_2)$ (i.e. noise alone) by taking the limit of $p(d|f_1)$ as the step edge amplitude A goes to zero.

The noise+step edge signal is assumed to be differentiated twice and filtered with a Gaussian lowpass filter with space constant σ_f. The autocorrelation function of a noise process obtained by passing white noise (with a variance σ_n) through such a filter is given by [31]

$$\psi(\tau) = \frac{\sigma_n^2 \sigma_f}{8\sqrt{\pi}} \left(3 - 3\left(\frac{\tau}{\sigma_f}\right)^2 + \frac{1}{4}\left(\frac{\tau}{\sigma_f}\right)^4\right) e^{-\frac{\tau^2}{4\sigma_f^2}}$$

where σ_n^2 is the noise variance. We will later need to know the value of some of the derivatives of the autocorrelation function evaluated at zero. We can perform a McLaurin series expansion to $e^{-\frac{\tau^2}{4\sigma_f^2}}$ and

see that the derivatives of the autocorrelation function evaluated at zero are given by

$$\psi^{(N)}(0) = \frac{\sigma_n^2(-1)^{\frac{N}{2}}(N+4)!}{\left(\frac{N}{2}+2\right)!2^{N+5}\sigma_f^{N-1}\sqrt{\pi}}$$

for N even, and is identically zero for odd N.

In determining an approximation to $p(d|f_1)$ we will assume that both the noise and the edge signals can be treated as linear near the location of the step edge (somewhat more restrictive assumptions were used in [94] in a similar application). The slope m_s of the filtered step edge will be (by looking at the step response of the Gaussian bandpass filter)

$$m_s = \frac{A}{\sigma_f^3\sqrt{2\pi}}$$

Let us assume that the true step edge is located at $x = 0$. Let us denote the slope of the noise signal by m_n (which will be a random variable) and assume that the value of the noise signal at $x = 0$ is n_0. Then the displacement of the signal edge (i.e. the zero crossing of the noise ramp and the signal ramp) is, as shown in figure 8.1, given by

$$d = -\left(\frac{m_s v\Delta t - n_0}{m_n + m_s}\right) = \frac{s}{r}$$

We can obtain the conditional probability density of d given f_2 from the above equation if we have expressions for the densities of $s = m_s v\Delta t - n_0$ and $r = m_n + m_s$. The random variable s has a Gaussian distribution with a mean of $b = m_s v\Delta t$. It has a variance given by $\psi(0)$, where ψ is the autocorrelation function given above. The random variable r is also Gaussian, with mean equal to $a = m_s$ and variance $-\psi^{(2)}(0)$.

The probability density of $d = \frac{s}{r}$ is given by the following marginal distribution

$$p(d) = \int_{-\infty}^{\infty} |r| p_r(r) p_s(rd) dr$$

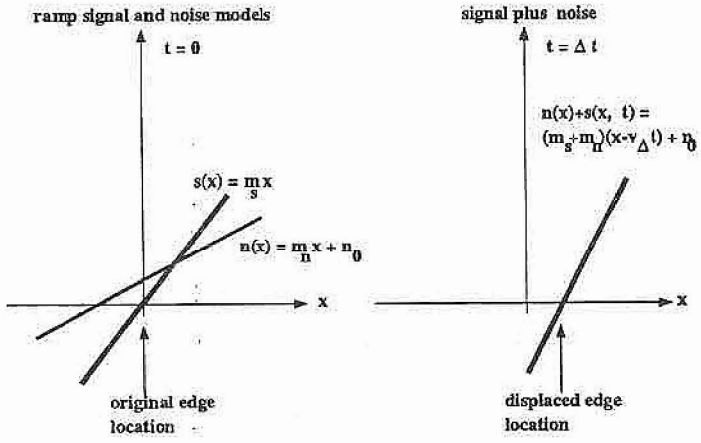


Figure 8.1: Computing the edge displacement using a ramp signal and noise model.

or

$$p(d) = \frac{1}{2\pi\sqrt{-\psi(0)\psi^{(2)}(0)}} \int_{-\infty}^{\infty} |r| e^{-\frac{(r-a)^2}{2\psi^{(2)}(0)}} e^{-\frac{(rd-b)^2}{2\psi(0)}} dr$$

One can expand out the terms in the exponentials and complete the square to yield

$$p(d) = \frac{e^{-\left(\gamma^2 - \frac{\beta^2}{\alpha^2}\right)}}{2\pi\sqrt{-\psi(0)\psi^{(2)}(0)}} \int_{-\infty}^{\infty} |r| e^{-\left(r-\frac{\beta}{\alpha^2}\right)^2 \alpha^2} dr$$

where

$$\alpha^2 = \left(\frac{1}{-2\psi^{(2)}(0)} + \frac{d^2}{2\psi(0)}\right)$$

$$\beta = \left(\frac{a}{2\psi^{(2)}(0)} + \frac{bd}{2\psi(0)}\right)$$

and

$$\gamma^2 = \left(\frac{a^2}{2\psi^{(2)}(0)} + \frac{b^2}{2\psi(0)}\right)$$

After some algebra and evaluating the integral, $p(d|f_1)$ can be shown to be equal to

$$p(d|f_1) = \frac{e^{-\left(\gamma^2 - \frac{\beta^2}{\alpha^2}\right)}}{2\pi\alpha^2\sqrt{\psi(0)\psi^{(2)}(0)}} \left[e^{-\frac{\beta^2}{\alpha^2}} + \left(\frac{\beta}{\alpha}\right)\sqrt{\pi}\,\mathrm{erf}\left(\frac{\beta}{\alpha}\right)\right]$$

We can obtain $p(d|f_2)$ from the above expression for $p(d|f_1)$ by taking the limit as $a \to 0$. This yields

$$p(d|f_2) = \frac{1}{2\pi\alpha^2\sqrt{\psi(0)\psi^{(2)}(0)}}$$

8.1.2 BAYESIAN EDGE DETECTION DECISION PROCESS

To obtain a solution to our problem of determining whether an edge is due to noise or not, one need only compute the values for the conditional probabilities $p(f_1|d_1, d_2, ..., d_N)$ and $p(f_2|d_1, d_2, ..., d_N)$ and take as the solution the f that corresponds to the maximum of these two. This decision rule comes from the application of the statistical decision theory described in chapter 2. Here we take as our decision rule that which minimizes the average risk, subject to the constant costs: $C_{1-\alpha} = C_{1-\beta} = 0$, $C_\alpha = C_\beta$. This was seen in chapter 3 to be a threshold rule: decide "edge" if $\Lambda < 1$, else decide "noise edge", where Λ is the generalized likelihood ratio

$$\Lambda = \left(\frac{p}{q}\right)\left(\frac{p(\vec{d}|f_1)}{p(\vec{d}|f_2)}\right) \qquad 8.1$$

where p is the *a priori* probability of an edge being a true edge and q is the *a priori* probability of an edge being a noise edge. Plots of the likelihood function (for $p = q$) are shown in figure 8.2 and 8.3.

For both graphs the value of the filter space constant is $\sigma_f = 8$. In figure 8.2 is shown the dependence of the likelihood function

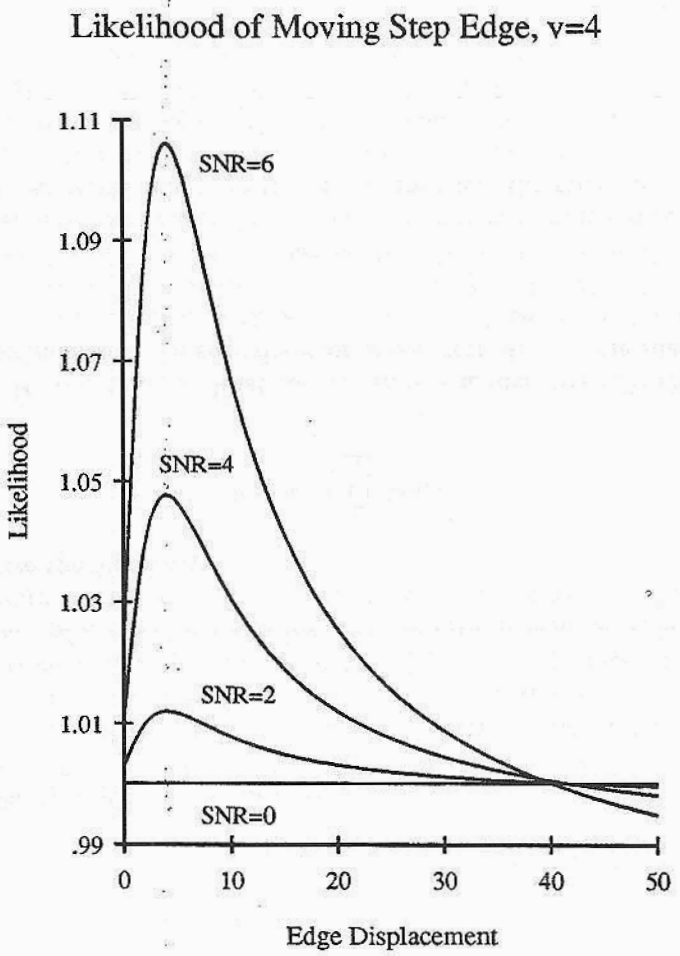


Figure 8.2: The likelihood as a function of the signal to noise ratio, for an edge velocity of 4 pixels/time step.

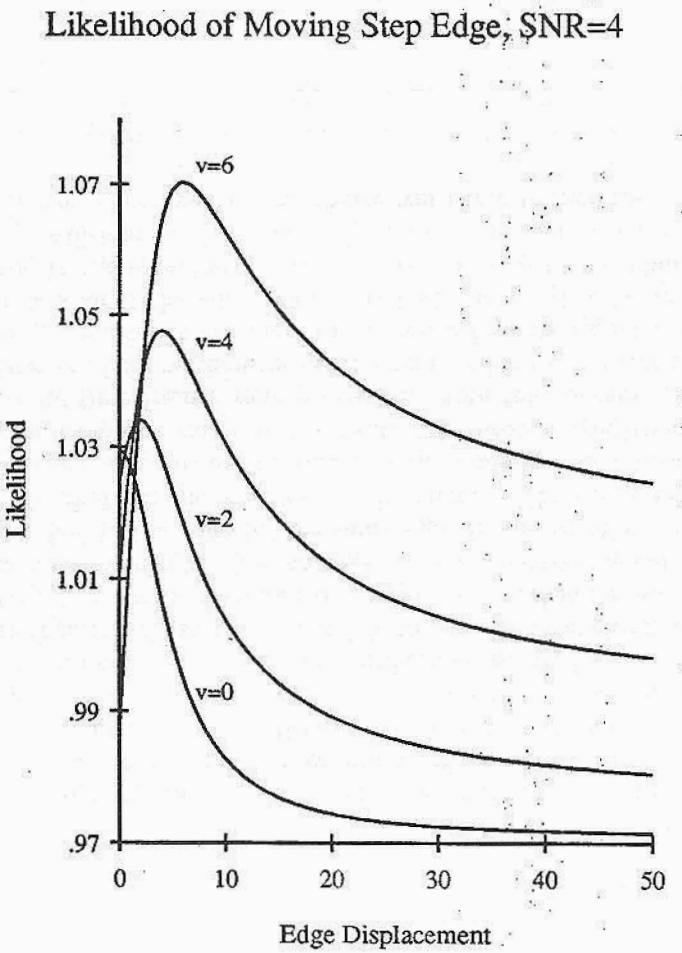


Figure 8.3: The likelihood as a function of the edge velocity, for a signal to noise ratio of 4.

on the signal to noise ratio (defined as $SNR = \frac{A}{\sigma_n}$), for an edge velocity of 4 units/time step. Note that the variance of the likelihood decreases as the SNR increases and the peak likelihood increases with increasing SNR. Figure 8.3 shows the dependence of the likelihood on the velocity of the step edge, for a SNR of 4. Note the increase in the peak likelihood with increasing velocity. Note also the increase in the mode of the likelihood distribution with increasing velocity.

The Bayesian approach also permits the computation of the reliability of the solution. This reliability measure is some monotonic function of the ratio between the two conditional probabilities, $P(f_1|d_1, d_2, ...)$ and $P(f_2|d_1, d_2, ...)$ (i.e. of Λ). If the value of this ratio is very high then the reliability of the decision, regarding the presence of a signal edge, is high. If the ratio is near zero then the reliability of the decision will be low. We will use as our reliability measure the following:

$$R = \log\left(\prod_{i=1}^{N} \Lambda(d_i)\right)$$

It is seen that, at least for our simple model, the reliability of the decision that the edge detector makes regarding a true edge (i.e. $f = f_1$) is a function of three parameters. These are (i) the signal to noise ratio (the greater the SNR the higher the reliability), (ii) the velocity of the edge (the larger the velocity, the more reliable is the decision for a finite N) and (iii) the number N of measurements taken. As more and more time steps are used, the reliability of the edge detection process increases. The dependence on the edge velocity arises due to the falloff in the probability that an edge is due to noise at large displacements. In practice the increase in reliability with velocity will diminish since for any reasonable temporal edge tracking procedure (such as nearest neighbor matching) edges will be improperly tracked at high velocities. If $v\Delta t$ is larger than the average distance between noises edges the true edges will likely be matched to nearby noise edges instead of the true edge. It would be interesting to see whether or not the results of the reliability analysis of this edge detection algorithm would be reflected in psychophysical experiments on detection of moving edges in noise by human observers. We would predict that, if the human visual system uses a temporal coherence method for detecting edges, the psychophysical experiments should indicate that moving edges are detected more quickly than slowly moving edges and that decreasing the signal to noise ratio increases the length of time required to detect an edge.

In figures 8.4 to 8.6 we show the results of a simple implementation of the temporal coherence edge detection algorithm described above. In this implementation candidate edges were detected using a zero crossing detector on the second derivative of the combined signal and noise waveform. Edges were detected and localized in the first time frame. The initial edge displacements, d_1^k, for each edge, k, were then initialized to zero, and N to one. The waveform at the next time frame was taken and the edges were tracked by matching to their nearest neighbor (i.e. an edge in the waveform taken at $t = 0$ was matched to the edge (of the same sign) in the waveform taken at $t = \Delta t$ that had the smallest displacement). The displacements d_2 were then taken to be the values of the difference in position of the edges from one frame to the next. The process continued until $N = 20$ time frames were processed. Note that because of the simplistic tracking mechanism used, a signal edge can be lost by erroneously matching it to a noise edge, especially if the signal to noise ratio is low or if the edge velocity is high (although this was not observed in the experiments we ran). In this case the reliability of the edge will start to fall. Similarly noise edges can be tracked to signal edges, whereupon the reliability will begin to increase.

In figure 8.4 we see the reliability as a function of time of all the edges (due to noise and the step) for a signal to noise ratio of 4, and a step edge velocity of zero. It is seen that the reliability of the step edge increases linearly with time, and is positive. The reliability for the noise edges, on the other hand decrease (more or less linearly) with time, and are negative. It is very easy to discriminate between the step edge and the noise edges based on the reliability value.

In figure 8.5 we see the reliability as a function of time of the

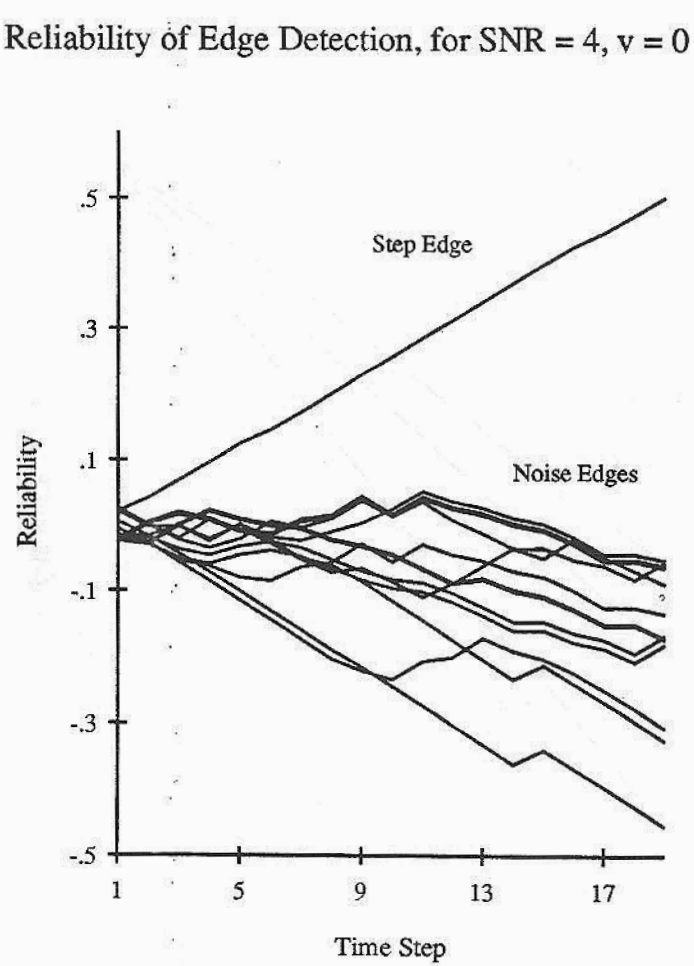


Figure 8.4: The reliability of all edges, for SNR = 4, and $v = 0$.

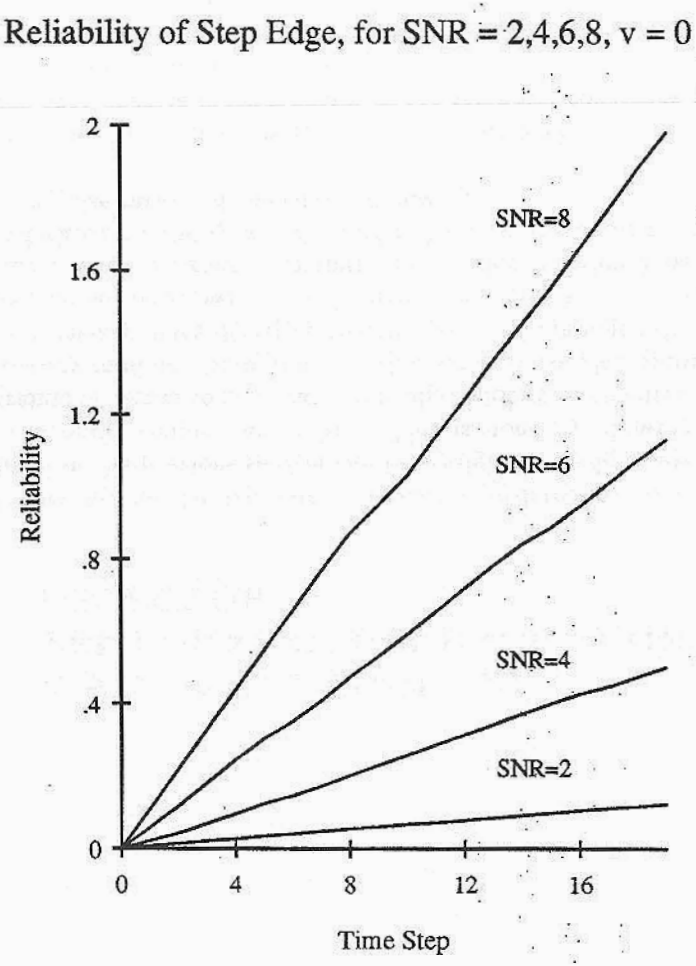


Figure 8.5: The reliability of the step edge, for SNR = 2,4,6 and 8, with $v = 0$.

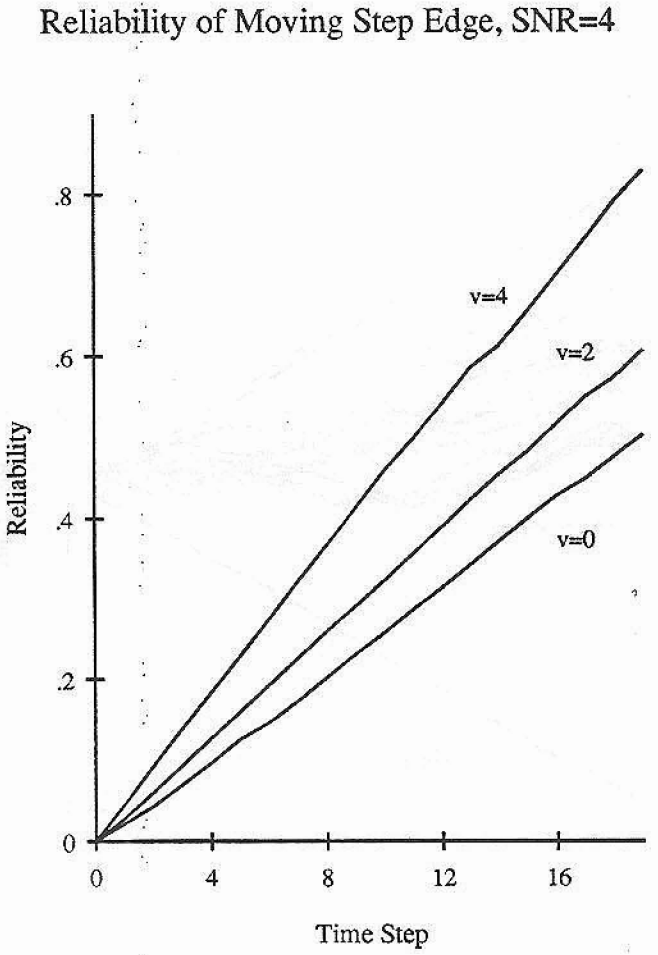


Figure 8.6: The reliability of a moving step edge, for SNR = 4, and v = 0,2, and 4 pixels/frame time.

step edge for a stationary edge, for four different values of the signal to noise ratio. It is seen that the slope of the reliability curve is an increasing function of the signal to noise ratio.

In figure 8.6 we we see the reliability as a function of time of the step edge for a moving edge, for four different values of the edge velocity. The signal to noise ratio is held at four for each case. It is seen that the slope of the reliability curve increases as the edge velocity increases.

8.2 A STRONGLY COUPLED TEMPORAL COHERENCE EDGE DETECTOR

The weakly coupled temporal coherence edge detection algorithm described in the previous section can be straightforwardly converted into a strongly coupled algorithm. This is done by allowing the constraints on signal to noise ratio and edge velocity to be adjustable by auxiliary modules. In practice, such a strongly coupled approach will be necessary, since the signal to noise ratio and the edge velocities will usually not be known *a priori* , indeed they may actually change over time. Thus, in order to compute the decision function λ, as well as the decision reliability, we will need to have estimates for both the signal to noise ratio and the edge velocity.

One could have modules for estimating the SNR and the edge velocity that operate independently of the temporal coherence edge detector module. Alternatively, one could use the information produced by the edge detection module in obtaining the SNR and velocity estimates. For example, the temporal coherence module, as part of the edge detection process, compiles statistics for the step edge displacement d (i.e. the distribution $p(d|f_1)$). We can see from figure 8.2 and equation 8.1 that the variance of d increases as the signal to noise ratio goes down (we assume that $p(d|f_2)$ is constant here).

Thus, we would expect that we could obtain an estimate for the SNR from the estimate of the variance of d. There are a couple of problems with this approach, however. The first is that we do not know *a priori* which edges are noise and which are signal edges. Since to obtain the desired variance requires us to have an estimate of $p(d|f_1)$ we must know which edges are due to the signal. The second problem is that the relationship between the signal to noise ratio and the variance of $p(d|f_1)$ is quite complex, and, in fact, the variance of $p(d|f_1)$ may be infinite. As the SNR goes to zero $p(d|f_1)$ approaches $p(d|f_2)$ which is Cauchy distributed, and hence has infinite variance. Measurements of the sample variance of $p(d|f_1)$ over time taken during the experiments described above show that this variance is poorly correlated with the signal to noise ratio. We conclude from this that it is probably not a good idea to estimate the SNR from the statistics of $p(d|f_1)$.

A possible strongly coupled approach to the SNR estimation problem would involve the following iterative process. First we assume that a randomly selected edge (or it could be the first in the signal) is the step edge. We then measure the magnitude of its slope and divide this by the average of the magnitude of the slopes of the rest of the edges. We compute this ratio for every candidate edge and associate the signal to noise ratio with the largest of these ratios. We then use this SNR in the edge detection procedure as usual. At subsequent time steps we recompute the SNR using as the candidate step edge that edge which has the largest (positive) reliability measure.

The estimation of the edge velocity is much more straightforward. Assuming that we use as the step edge that which had the highest SNR in the SNR estimation process, we can take as the edge velocity the average of the displacement of this edge per time step. At subsequent time steps we use as our measure of velocity the displacement of the edge with the highest reliability. This will allow us to lock onto the proper edge if we do not have it initially. Figure 8.7 shows the variation of the estimate of the mean step edge displacement over time for one of the experiments described in the previous section. It is seen that this estimate approaches the value of $v\Delta t$, where v

is the velocity of the step edge. Thus, we expect that the mean of the step edge displacement can be used to compute the value of v in our computation of the edge reliability. This method of estimating the step edge velocity has difficulty when the velocity changes over time. In such a case the mean of the step edge displacement will not be a good estimate of the instantaneous edge velocity. An improved estimate can be obtained by weighting the displacement measurements in determining the mean, so as to emphasize recent displacements over older displacements. There is a tradeoff between the amount of blurring of the velocity value due to excessive filtering and the amount of error in the estimate of the mean (which decreases as more samples are taken). In order to come up with an appropriate temporal filter one must specify a model for the temporal variation of the velocity, and then optimize the filter, taking this model into account, with respect to a suitable optimality criterion. Hwang and Clark [70] go through this process of specifying a statistical model for the temporal velocity variation and use an approach similar to that of Canny [28] to define an optimality criterion and produce the optimal temporal filter. We could use this filter in estimating the mean of the step edge displacement in our application.

The strongly coupled temporal coherence edge detection method described above is a recurrent one since we are feeding back the information about the reliability of a candidate edge to the velocity and signal to noise ratio estimation modules. These SNR and velocity estimates are then used to update the reliability measures. Like all recurrent strongly coupled methods one has to be concerned about the convergence of the algorithm. If the wrong edge is used in the estimation of the SNR and velocity then the reliability of the true edge may not become large enough to fix the error. This is expected to be a problem at low SNR values, which is, unfortunately, precisely where the application of the temporal coherence is intended to be most useful.

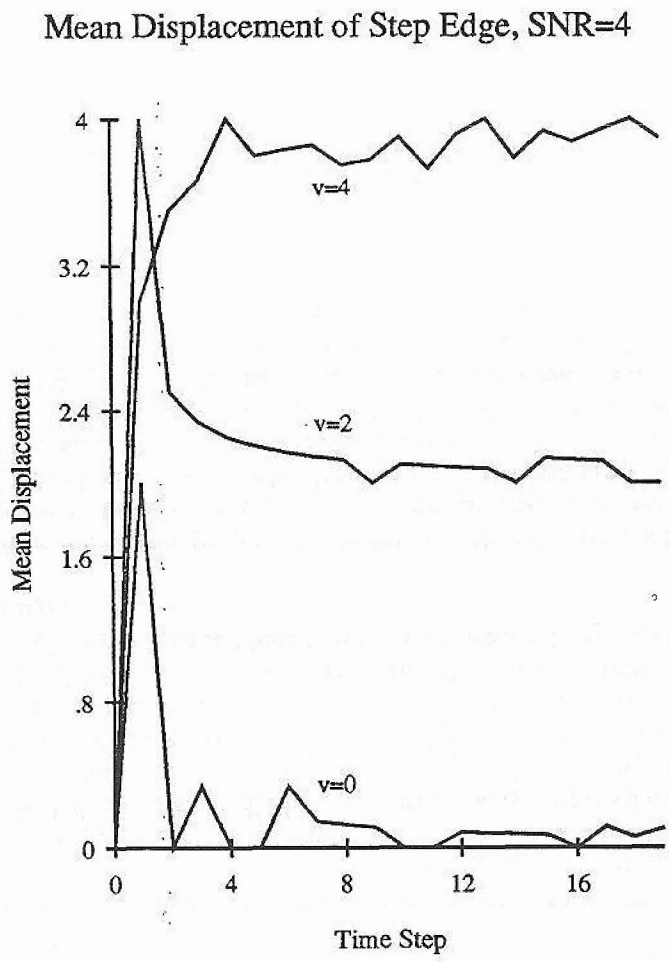


Figure 8.7: The mean displacement of the step edge for SNR = 4, and v = 0,2, and 4 pixels/frame time.

8.3 TEMPORAL SAMPLING

Let us begin our brief study of the temporal aspects of data fusion by looking at the temporal sampling question. Consider, to start with, a particularly simple world that can be described by a time varying parameter vector $\vec{f}$. Suppose we want to know the value of $\vec{f}$ for all points in time (i.e. specify $\vec{f}(t)$). In any practical system we cannot continuously measure and process data, so we have to sample in time. Let us denote the (infinite) set of the samples by $\{\vec{d}\}$. Let us, for the purposes of argument relate the sample set to $\vec{f}(t)$ through the following simple image formation model:

$$\vec{d}_k = \vec{f}(t_k) + \vec{n}_k$$

where $\vec{d}_k$ is the k^{th} sample in the set $\{\vec{d}\}$, and $\vec{n}_k$ are Gaussian distributed random variables, assumed to be uncorrelated with each other (the components of the $\vec{n}_k$ are assumed to be uncorrelated as well). The t_k are the points in time that samples are taken. The image formation model is seen to be (modulo a normalization factor)

$$p(\{\vec{d}\}|\vec{f}(t)) = \prod_k e^{-\frac{1}{2}(d_k - f(t_k))^T M_k^{-1})(d_k - f(t_k))}$$

where M_k is the covariance matrix of $\vec{n}_k$. If we apply the Bayesian formulation, the problem of estimating $\vec{f}(t)$ from the sample set $\{\vec{d}\}$ obviously depends on the prior model we use to describe constraints on $\vec{f}(t)$. Since we have only a single $\vec{f}(t)$ we need only worry about modeling the relationship between the components of $\vec{f}(t)$ and the temporal properties of $\vec{f}(t)$. It should be obvious that, if we have a uniform prior model, there is no unique solution for $\vec{f}(t)$ to be obtained from the data $\{\vec{d}\}$. There are a infinite number of possible functions $\vec{f}(t)$ that have the samples $\{\vec{d}\}$, even in the noise free case. Thus we must specify a nontrivial prior model in order to constraint the solution sufficiently. To examine this problem more closely let us restrict our self to the one dimensional case, where $f(t)$ is a scalar function of time. The approach detailed below can be extended to the multidimensional case, but is more difficult and restrictive.

A commonly used constraint on functions is that they can be represented in terms of an integral transform [82] as follows:

$$f(t) = \int_I K(x,t)g(x)dx$$

where $g(x)$, and $K(x,t)$ for all real t, are in $L_2(I)$ (i.e. are square integrable over the interval I). An example of such a constraint is the bandlimitedness constraint, which is obtained if we set $K(x,t) = e^{jxt}$. The integral transform in this case is the Fourier transform. For non-Fourier kernels we no longer have a bandlimit constraint *per se*, [32, 141], but the constraint is similar to a bandlimit constraint, so we will call it a *generalized bandlimit constraint.*

The generalized bandlimit constraint can be represented in a Bayesian implementation as

$$p(f) = e^{-\int_{x \neq I} |g(x)|^2 dx}$$

where $g(x)$ is obtained by inverting the integral transform (which is assumed to be well posed). The *a priori* probability density given above is maximized by $g(x)$ having zero energy outside the interval I. This occurs only if $f(t)$ satisfies the generalized bandlimit constraint. If $f(t)$ does not satisfy this constraint $g(x)$ will have non-zero energy outside the interval I and the *a priori* probability will be less than the maximum.

Let us define a family, indexed by k, of generalized bandlimited signals (i.e. that are good candidates to be $f(t)$ before we take the data into account) as follows:

$$S_k(t) = S(t,t_k) = \frac{\int_I K(x,t)K^*(x,t_k)dx}{\int_I |K(x,t_k)|^2 dx}$$

where K^* is the complex conjugate of K. The members of this family obviously satisfy the generalized bandlimit constraint, since the denominator is a constant, and the $K^*(x,t_k)$ term in the numerator corresponds to $g(x)$ in the definition of the generalized bandlimit constraint (and since $K(x,t) \in L_2(I)$ for all real t so is $K^*(x,t_k)$).

It was shown by Kramer [75, 82] that if there exists a countable set $E = \{t_k\}$ such that $\{K(x,t_k)\}$ is a complete orthogonal set on I then we can uniquely represent any function $f(t)$ that satisfies the generalized bandlimit constraint as a weighted sum of the $S_k(t)$ functions. That is,

$$f(t) = \lim_{N \to \infty} \sum_{|k|<N} f(t_k)S_k(t)$$

When we take the data into account, the optimal estimate of $f(t)$, i.e. the one which maximizes $p(f|\{d_k\})$, is given by

$$\hat{f}(t) = \sum_{k=-\infty}^{\infty} d_k S_k(t)$$

Let us summarize the steps given above for finding $f(t)$ from the sampled data $\{d\}$. First we must determine a suitable *a priori* constraint for $f(t)$. This is done by specifying the kernel $K(x,t)$ of the generalized bandlimit constraint. If we are given a sample sequence $\{t_k\}$, we must ensure that our $K(x,t)$ is such that $\{K(x,t_k)\}$ forms a complete orthonormal set on I. If it does not, then we will not be able to uniquely determine $f(t)$ from the data and additional constraints will be required. Alternatively, we may choose a particular $K(x,t)$ and determine a suitable sample sequence $\{t_k\}$. In the case of the Fourier kernel, $K(x,t) = e^{-jxt}$, the problem of determining allowable sample sets $\{t_k\}$ has been studied by Paley and Wiener [118], Levinson [88] and others. The conclusion (so far) is that, for $\{e^{-jxt_k}\}$ to be a complete orthogonal set, the set $\{t_k\}$ must be uniform ($t_k = kT$), or almost so (with a deviation from uniformity that does not deviate by more than about 25 %). Once $K(x,t)$ and $\{t_k\}$ have been specified we create a family of functions which depend on $\{t_k\}$ and which satsify the generalized bandlimit constraint. This set of functions actually form a basis for the set of all functions which satisfy the generalized bandlimit constraint. We can thus represent $f(t)$ by a weighted sum of these functions. The weights, for the Bayesian MAP estimate of $f(t)$, correspond to the data samples d_k. If we take the Fourier kernel, with uniform sampling, we get the standard Shannon [140] sampling theorem.

The important point to be gained from the above discussion is that by applying a suitable constraint to $f(t)$ we can obtain both a unique solution to the problem of determining $f(t)$ from the data, and information as to when we should take our measurements. We saw that the specification of the sample times comes from the type of constraint that we imposed on $f(t)$. As with any *a priori* constraint, if it is invalid, the resulting solution that we obtain with the above method will be incorrect. The error that arises in applying the wrong constraint in the above problem is usually referred to as aliasing error.

8.3.1 COMPUTATIONAL CONSTRAINTS

In discussing the above problem we have been consider quite a simple case. In practice, determining a suitable constraint may be quite difficult. We will almost always end up by imposing a constraint which is invalid, and hoping that the resulting error is small. What constraint we choose is more often than not driven by characteristics of the system that is performing the computations rather than by an analysis or modeling of the environment. The most fundamental limitation that we are faced with is that estimates of world parameters must be periodically obtained. We cannot wait forever to get an infinite sample set, we must perform our estimates using temporally local operations. This will introduce errors, usually refered to as truncation error, in estimating $f(t)$. Another important limitation arises due to the fact that computations cannot not be performed instantaneously. For example, a piece of computer hardware that we are using in a robotic system may be able to do a particular estimation process once every T seconds. Thus our sampling rate cannot be faster that once every T seconds, without having computations piling up. Thus we can expect that if the world parameters are changing too rapidly we will misperceive them due to our insufficient sampling rate. Another way of looking at this is to say that our hardware limited sample sequence arbitrarily imposes a constraint on the world parameters. If this constraint is invalid, then our perception of the world will be incorrect.

When we consider the fusion of data coming from two sensory modules in time we must be concerned with the temporal characteristics of these modules. The uncertainty of a piece of information is usually a function of the computation, or sampling, rate of the module that produces it. For some algorithms, the longer it runs the more accurate it is. Also, given two algorithms, the one that takes longer is generally the more accurate of the two (although exceptions are certainly possible due to inefficiencies in the implementation of a given algorithm). Thus a source may be twice as accurate as another but only give results half as often. Which source is more useful? From the point of view of Shannon's sampling theory (and also the more general theory due to Kramer discussed earlier) the answer would depend on the "bandwidth" of the information that is to be measured. If the bandwidth is very low then one will not lose anything by sampling at a low rate, so that the module with the high accuracy would be preferred. If, however, the bandwidth of the information was high, then the slow sampling rate of the more accurate source would result in a greater amount of aliasing error than the module with the faster sampling rate. As an example of this principle, consider the fusion, by weighted summation, of two data sources, x_1 and x_2 to yield $\hat{x}$:

$$\hat{x} = \alpha x_1 + (1 - \alpha)x_2$$

where α is a weighting value between 0 and 1. The value of α, following the principles of weak coupling expounded on in Chapter 4, should be a function of the relatively reliability between x_1 and x_2. For example, $\alpha = \frac{1}{1+R}$, where R is the reliability of x_2 divided by the reliability of x_1. We would then modify this to take into account the effects of the sampling rate and computational complexity on the reliability of the modules.

Other implementation related issues that one must be concerned with include the effects of computational latency, and of data obsolescence. The latency issue arises when one has pipelined computation wherein the rate, or throughput, of computation is quite high, but

where there is a delay between the input of the data and the computation of desired quantities from the data. This is a problem in active systems where the sensory data is used to tell the system how too interact with the environment. If the latency is too large then the derived information is obsolete and may lead to inappropriate actions by the system. The latency is also a problem when performing fusion of data from modules that different latencies. This is especially true in recurrent strong coupling due to the complicated dynamics that are involved.

As researchers explore "active" sensory information processing systems, where time plays an important role, the need for accurate modeling of the temporal characteristics of both the environment and of the system itself will become increasingly important.

8.4 ACTIVE DETERMINATION OF CONSTRAINTS

As mentioned in chapter 1, one is often faced with the situation wherein one or more of the constraints used in a sensory module is invalid. One would ideally like to be able to detect when a particular constraint is invalid and replace it in the algorithm with a more suitable constraint. It is very difficult, however, to determine the validity of a given constraint in a static situation using the data from a single sensor, at a single instant of time. If one has multiple sensors, then they can *fuse* their information to produce check the validity of their associated assumptions and constraints. If the constraints are judged invalid they can be altered and a new solution determined. This would lead to a strongly coupled data fusion approach.

The constraint adaption process promises a way out of the dilemma that faces the designer of a Bayesian information processing system. The dilemma is that adding more constraints to a problem eases its solution, while increasing the likelihood of obtaining an incorrect solution. The problem arises since we usually do not know beforehand which constraints are going to be valid in a given situation, so to make our algorithms as generally applicable as possible, we must weaken our constraints. If we do this, however, the problem may become insoluble, or its solution impractical, as we haven't applied enough constraint. Since most designers of sensory systems would rather get possibly wrong solutions than no solutions at all, they typically add enough "reasonable" constraints to the problem to allow solutions to be obtained. Instead of adding in possibly invalid and probably superfluous constraints the constraint adaption process allows us to begin with a small number of weak general constraints and alter these constraints based on measurements taken by various sensory modules.

Another approach is to use the additional information obtained from the same module over time, as the sensors move through space. The consistency of the output of a sensory processing module over time is a measure for how valid the constraints embedded in the module are. If the wrong constraints are being applied in the processing of the module, then in general, the results of the module at two different times will be inconsistent. Based on the form of the inconsistency, the constraints can be adapted in a way to make subsequent solutions more consistent as the sensors move. If invalid constraints are used in a static sensory processing module, the result will be incorrect. The task of validating the result in the absence of additional data is very difficult. We have seen in earlier chapters that information from other modules can be used to validate the results of a given sensory processing module.

As an example of the temporal inconsistency of a vision algorithm, many shape from shading algorithms provide estimates of object shape that are overly flat, or otherwise excessively warped by whatever smoothness constraints are used by the algorithm (c.f. the discussion on the shape from shading algorithm in chapter 7). In addition there will occur errors in shape due to factors other than improper constraints (e.g. noise in the image, finite precision in computation, and so forth) which will tend to be independent in time.

Thus, as one moves the imaging sensor and repeats the shape from shading computation, the resulting surface, related to some fixed reference frame, will typically be different than the one computed in the previous time step. The perception will therefore be of a surface that is deforming in time or moving non-rigidly. However, such a perception is an impossibility if we assume that only the observer (the camera) is moving and the object is stationary (and not deforming, pulsating, or whatever). Thus one can conclude that the stationary object or rigidity constraints are *inconsistent* with the type of smoothness constraint utilized in the shape from shading algorithm.

Armed with the knowledge of the inconsistency of the shape from shading smoothness constraint with the more likely object stationarity constraint, we can attempt to modify the smoothness constraint in a manner that reduces the inconsistency. It is not immediately obvious how this might be done, however. One might consider replacing the smoothness constraint with a consistency constraint. That is, instead of determining the surface whose reflectance is closest to that of the observed reflectance subject to maximization of a smoothness measure, we look for a surface whose reflectance is closest to that of the observed reflectance and which undergoes a minimum of deformation over time. Under closer inspection, however, it is clear that such temporal consistency constraints are too weak to be applied directly in such a fashion. This is due to the fact that inappropriate solutions can be temporally consistent just as easily as reasonable solutions. The consistency constraint has no power to distinguish good solutions from bad, only temporally consistent solutions from inconsistent solutions.

The consistency constraint may best be thought of as a meta-constraint, one which "constrains the constraints". For example, as noted above, we cannot apply the temporal consistency constraint directly, however we can use it to constrain the form of the smoothness constraint that we do apply directly. In order to apply the consistency meta-constraint to the smoothness constraint it is necessary to adaptively change the form of the smoothness constraint as a function of measures of temporal consistency. It is evident that

this process implies a form of strongly coupled data fusion, where a module measuring temporal consistency is used to alter the *a priori* constraints used in the shape from shading process. The data fusion could be referred to as *active data fusion*, as the alteration of the *a priori* constraints is done as a result of the motion of sensing system. This motion can be purposive, and controlled so as to yield maximum expected inconsistencies. That is, if we have a choice regarding which direction we are to move our sensors, we should choose this direction so as to maximize the information concerning the validity of our *a priori* constraints. This is accomplished by moving in the direction that we expect will result in the largest inconsistency between computed surfaces over time. There is a distinction to be made between these active constraint adaption algorithms and the standard active vision methods [2] along the lines of those described in chapter 4. The active data fusion process will necessarily be a strongly coupled process since the motions produced will be dependent on the outputs of the sensory modules and will in turn affect the data that is input to the sensory modules (via alteration of the prior and image formation models).

One must be careful to distinguish between inconsistencies that result from improper *a priori* constraints from inconsistencies that arise from errors or noise in the raw data itself. These may be distinguishable in some cases by the statistics of the spatial and temporal distribution of the inconsistencies. It will often be the case that noise induced inconsistencies will be uncorrelated both temporally and spatially, and therefore have fairly significant high spatial and temporal frequency components, while the inconsistencies due to improper constraints will tend to be correlated over time and space and contain mainly low frequency components.

As an example, consider the regularized version of the shape from shading problem given in [23], expressed as an energy functional minimization problem, where the shape (described by the surface normals $\hat{n}(x,y)$) is that which minimizes the following functional:

$$\int_A [(E(x,y) - \hat{n} \cdot \vec{s})^2 + \lambda(||\nabla \hat{n}(x,y)||^2) + \mu(x,y)(||\hat{n}||^2 - 1)]dA$$

where the $\mu(x,y)$ is a Lagrangian multiplier function that enforces the constraint that $\hat{n}$ be a unit vector. The second term in the energy functional imposes a smoothness constraint ($||\nabla\hat{n}(x,y)||^2 = ||\frac{d\hat{n}}{dx}||^2 + ||\frac{d\hat{n}}{dy}||^2$), and the parameter λ sets how much smoothing is applied. As pointed out in [64] the solution to this minimization problem for nonzero λ is "flattened" due to the effect of the smoothness constraint. This "flattening" is a result of the fact that the particular smoothness constraint used is not really valid (even when the surface is "smooth" according to some other definition of smoothness). If we move the camera and repeat the energy functional minimization we will get a new shape estimate, which will again be "flattened". In this case, however, the direction of the flattening will be different than in the first computation. Thus, the object, whose shape is computed using the above technique, will appear to deform as we move the camera. This deformation indicates an inconsistency in the solutions implying the invalidity of our smoothness constraint.

One should therefore arrange the energy functional so that the temporal inconsistency of the shape estimation process is minimized. That is we need a temporal consistency constraint. The simplest way to do this is to force the surface normal maps to be the same for the two images (which is another way of stating the temporal consistency constraint). Thus, we can minimize the following functional:

$$\int_{A_1\cap A_2} [(E(x,y,t) - \hat{n}\cdot\vec{s})^2 + \mu(x,y)(||\hat{n}||^2 - 1) + \lambda[(E(x,y,t+\Delta t) - T\hat{n}\cdot\vec{s})^2]dA$$

where $T()$ represents the coordinate transformation between the surface as viewed at one time (A_1) and the surface viewed at another (A_2), and depends on the (known) motion of the camera (in general this will be a function of the height of the surface, but not if orthographic projection is assumed). Note that two images are enough to uniquely define the surface normal map, and that the smoothness constraint is not needed for regularization purposes (as the two images are sufficient to regularize the problem).

Another approach is to use an object rigidity constraint as our temporal consistency constraint. We expect that the object will remain rigid as we move, and not deform, so that the surface normal maps, relative to the reference frame, obtained at two different times should be the same. This leads to a different sort of energy functional to be minimized:

$$\int_{A_1\cap A_2} \left[(E(x,y,t) - \hat{n}\cdot\vec{s})^2 + \mu(x,y)(||\hat{n}||^2 - 1) + \lambda||\frac{d\hat{n}}{dt}||^2\right] dA$$

The time derivative can be approximated by a difference

$$\frac{d\hat{n}}{dt} \approx Tn(x,y,t+\Delta t) - n(x,y,t)$$

In this approach the estimated surface shape will initially be flattened as in the single view shape from shading algorithm, but will become less flattened as time goes on and the rigidity constraint takes effect.

The examples described above illustrates the application of a temporal consistency constraint in place of a surface smoothness constraint. A more general approach is to modify the smoothness constraint based on the temporal consistency in a manner that effectively enforces a temporal consistency constraint. In order to do this we need to specify a parametrized space of smoothing operators, and then adjust the parameters using gradient descent on a consistency measure. For example we could specify a general quadratic first order smoothness constraint with:

$$S(p,q) = \frac{1}{(1+p^2+q^2)^2}(P_1p_x^2 + P_2q_x^2 + P_3p_y^2 + P_4q_y^2 + P_5p_xp_y + P_6p_xq_x + P_7p_xq_y + P_8p_yq_x + P_9p_yq_y + P_{10}q_xq_y + P_{11}p_x + P_{12}p_y + P_{13}q_x + P_{14}q_y)$$

where the $P_i(p,q)$ are second order polynomials in p and q. The smoothness constraints are thus parametrized by $6\times 14 = 84$ parameters. The smoothing operator used in the single view shape from shading algorithm is obtained with $P_1 = P_3 = (1+q^2), P_2 = P_4 = (1+p^2), P_6 = P_9 = -2pq, P_5 = P_7 = P_8 = P_{10} = P_{11} = P_{12} = P_{13} = P_{14} = 0$. We could use this as our starting point and modify the P parameter vector by performing gradient descent on the

(in)consistency measure $C = |\frac{d\hat{n}}{dt}|$. That is, alter the parameters of the smoothness constraint as follows:

$$\frac{\partial P_i}{\partial t} = -\frac{\partial C}{\partial P_i}$$

where P_i is one of the parameters of the smoothness constraint. It is conjectured that this will converge to $P_{12} = -P_{13} = 1$ with all other $P_i = 0$, resulting in an integrability type of smoothness constraint, which is known not to distort surfaces [64]. In general, however, if we do not know the shape of the object beforehand, the dependence of C on the parameters P_i cannot be precomputed. Instead this dependence must be estimated through an exploration in the space of P values.

Determining an efficient, and convergent method for deriving $\frac{\partial C}{\partial P_i}$ will form a major part of the development of the temporal constraint adaption method. A natural way of characterizing the suitability of a given smoothness constraint is to look at the space of minimal surfaces (i.e. the surfaces created by the energy function minimization as λ goes to infinity), compared to the actual surfaces being reconstructed. If the surface to be reconstructed lies within the space of minimal surfaces then the smoothness constraint is valid, otherwise it is invalid. One could try, then, to determine a smoothness operator that creates a space of minimal surfaces that is as large as possible (e.g. the integrability constraint) or one can try to adapt a smoothness constraint based on partial (actively obtained) information regarding the actual surface shape to adjust the smoothness operator to bring its space of minimal surfaces more into line with the current estimate of the surface. Techniques related to generalized cross-validation techniques [47] may possibly be effectively applied to this problem of adapting the parameters of the smoothing constraint.

Another example of the application of active data fusion arises in the structure-from-motion process. In this case, the observer is stationary and the object moves. It is known [151] that the optical flow field measured by the observer provides information regarding the shape of the object if we assume that the object is rigid. The optical flow itself, however, is often computed based on algorithms which

rely on possibly invalid assumptions, or constraints, much in the same fashion as the shape from shading algorithm described above. Thus the optic flow vectors may be incorrect, yielding, even with the application of the rigid body constraint, a shape or object structure that is inconsistent from time frame to time frame. That is, even though object rigidity was used as a constraint in the extraction of the object shape, the object can be perceived of as being non-rigid. This is because the rigidity constraint is applied only to produce object structure from optic flow fields at a single time step, while the non-rigid perception arises due to the invalid assumptions used in the optic flow measurement process.

Using the active temporal data fusion approach described above for the case of shape from shading one can develop a structure from motion algorithm that can produce object structures that are consistent over time with a rigidity assumption. This algorithm would measure the inconsistency (non-rigidity) between the object structure determined in two (or more) successive time frames and use this information to modify the constraints used in the optic flow field extraction process (and possibly the structure from optic flow process) in a way so as to reduce the inconsistencies. In this application of active data fusion we do not have the liberty of altering the object motion, so we have no control over the inconsistencies. For example, the object may move in a direction such that the inconsistencies are quite small. In this case the constraints will not be adapted very much, and would remain invalid.

Often one has the ability to alter certain parameters of the image formation model in a controllable fashion. In this way one could optimize the constraint adaption process or even the world parameter estimation process. For example, one could change the focal length of camera in order to obtain sharper focus, allowable more reliable data to be obtained. Or the alteration could be used to obtain independent information such as in depth from defocus algorithms [69, 120].

There has not been much work done to date on systems using the above constraint adaption methodology. One exception is the

work described by Krotkov and Kories [83]. They have constructed a strongly coupled approach to fusing depth from binocular stereo and depth from focus. In addition to providing a method for fusing the two depth modules, they define a statistical measure of consistency between the outputs of the two modules. This consistency measure is used by them to reject or accept the outputs of the modules, but could conceivably be used to adapt the constraints used in the activities of the modules.

The idea of adapting sensory information processing algorithms based on inconsistency information was suggested (but not attempted) by Krotkov and Kories [83] where they write:

> "It is also possible to feed back the information that cross-checking failed or could not be completed, and to reinitiate verification sensing *with different parameters.*"

Krotkov and Kories treat inconsistent measures of depth from the focus and stereo modules as an indication that the two modules are measuring the depth of two different physical quantities. If one took the point of view espoused in this book, they would instead consider the inconsistency to mean that one or more of the constraints implicit to the operation of the modules were invalid. In the particular case of the system of Krotkov and Kories, the invalid constraint could be the assumption that the depth being computed by the focus and stereo modules was in fact the depth of the same physical event. While this may be true for the bulk of the cases in which inconsistencies are obtained, there may be some cases for which other constraints are in fact being violated. For example, the inconsistency may be due to mismatching of two different features on the same object in the stereo module, rather than due to matching features of two different objects. The mismatching will typically result from the violation of some (usually implicitly assumed) constraint in the stereo module, such as smoothness of the disparity field, feature ordering, feature density, etc.

8.5 SUMMARY

- The information processing systems of organisms or machines that interact actively with their environment need to be concerned with the temporal properties of the structures in the environment.
- In the Bayesian formulation of information processing the temporal aspect of the environment is handled by specifying temporal constraints on the configuration of the world. We described a way to embed a generalized bandlimit constraint on world parameters, and how this leads to an optimal Bayesian estimate of the world parameter.
- The importance of modeling the temporal aspects of the sensory system, in terms of computational throughput, rate, and latencies was introduced. These aspects in practical systems tend to control the form of the constraints imposed on the world parameters rather than any models of the world.
- Faced with the prospect of not being able to specify the right constraints for a given environment, we would like a way to "learn" these constraints. We present a methodology for doing this based on the use of consistency measures, or meta constraints, obtained from the outputs of independent sensory modules, to adapt the prior constraints used by the modules in a way which maximizes the consistency measures.
- A powerful constraint natural temporal constraint is that of temporal coherence. This constraint implies that certain world parameters change slowly and smoothly, so that the values of these parameters at closely spaced time steps are closely related. We present an edge detection algorithm based on the application of this constraint. It uses an optimal Bayesian binary decision method to decide whether a candidate edge is due to noise or to an actual step intensity signal.

Chapter 9

Towards a Constraint Based Theory of Sensory Data Fusion

We have observed that information acquisition is required by organisms in order to determine the state of the environment that they are operating in. What information is required is related to the activities that the organism undertakes. In order to obtain the required world information one must invert the world-image mapping. In general this mapping is non-invertible so that extra information must be added to constrain the space of world configurations sufficiently to allow a unique solution to be obtained.

The extra information required to uniquely invert the world-image mapping is provided in sensory information processing systems in the form of physical, natural, and artificial constraints. Physical constraints are derived from studies of the physical and mathematical laws underlying the world. Natural constraints are contingent on the particular restricted domain the organism is expected to function in. These constraints are not guaranteed, nor expected, to be uni-

versally valid. Artificial constraints are a form of natural constraint wherein the expectations are at a higher cognitive level. Artificial constraints are even less likely than natural constraints to be valid in an arbitrary situation.

The determination of suitable constraints involves a characterization or modeling of the world and of the world-image mapping. This is typically a scientific process. Once the constraints have been determined one must embed them into a suitable algorithm. Bayes rule provides an intuitively satisfying means of embedding constraints into an information processing task. The principle aspect of this approach is the probabilistic representation of constraints. Solutions to the world-image mapping are assigned probabilities corresponding to their likelihood with respect to the constraints. In our view, the Bayesian formalism has two primary components, the image formation model and the prior model. The image formation model represents the conditional probability of a given data value arising from a particular world configuration. It captures the process by which data is generated from different world configurations. The prior model represents the *a priori* probability of the particular world configuration, and corresponds to the expectations of which configurations structures in the world are likely to occur.

Smooth energy function minimization is a special case of the Bayesian formulation which is well suited for imposing smoothness constraints. A suitable choice of the smoothness operator makes the solutions linear combinations of basis functions. By defining a Gibb's distribution one can give a probabilistic interpretation to minimizing energy functions. This shows that energy functions are a special case of the Bayesian formalism corresponding to Markov Random Field distributions. Markov random fields can be implemented with binary "line process" fields, which results in a special case of the Bayesian approach that allows the "breaking" of smoothness constraints at places where the smoothness constraint is inappropriate. We can generalize Markov random fields by adding in binary valued "Matching Element" fields. These matching fields allow us to specify correspondences between features in separate images, such as is required in long-range motion analysis or stereopsis.

For many vision problems, in particular those involving correspondence, such as motion analysis or binocular stereo, there are global constraints on the field which need to be satisfied. Statistical techniques give ways of imposing these constraints absolutely, unlike the more traditional weak methods of adding terms to the energy function which merely "encourage" the constraints to be satisfied.

Natural or artificial prior constraints may be invalid in certain situations. In these cases one would like to reduce the dependence of a sensory information processing algorithm on these constraints if possible. One way of doing this is to use information from independent sensory information processing modules as constraints on a given module. This data fusion, then, is a means for reducing dependence of possibly invalid *a priori* constraints. This is to be contrasted with the usual role of data fusion where the goal is to reduce the uncertainty (but not necessarily the validity!) in the estimate of a world parameter.

The principal applications of data fusion are: reducing uncertainty, getting enough information to provide a unique solution, and modifying possibly invalid prior constraints. Our emphasis in this book has been on the latter two applications. We classify fusional methods as being either weakly or strongly coupled, where the distinction is made by considering whether or not the operation of the modules are independent of each other. Weakly coupled fusion is characterized by the combination of the outputs of independent information sources. Each module operates independently of the others (although the modules may themselves perform fusional operations).

We divide the class of weakly coupled algorithms into a number of subclasses. The distinctions between these subclasses of weakly coupled fusion methods serve to point out that there are different reasons for performing sensory fusion and that different methods are called for in each case. The methods in class I represent the most common applications of data fusion and are used when the goal is

duce the uncertainty in a desired value, or when one wishes to reduce reliance on the assumptions of a given sensory module. Thes methods are characterized by the combination of independent sources of data, each of which is sufficient for estimating a given parameter.

The methods in class II are indicated when the modules available, taken individually, do not provide unique solutions. The methods in this class include the large group of algebraic methods, wherein the desired parameter can be computed from an analytic expression involving a number of independent variables, obtained from the independent sources of data. It was found the performance of many algebraic approaches was poor due to high sensitivity of the solution to errors in the data, as well as due to the ill-posedness resulting from the world-image mapping not being onto (that is, there are some image configurations that do not have any associated world configurations). Thus noise in the image data may cause there to be no solution. This was observed in the shape from shading application described in chapter 7.

The methods in class III are a combination of the first two classes of weakly coupled methods and are used when one wishes to combine the outputs of sensory modules that, by themselves, do not provide unique or stable outputs, and at the same time weight the information from the component modules according to their relative reliabilities in order to minimize the uncertainty in the fused output.

The second primary class of fusional algorithms, in our view, are those concerned with the adaption of the constraints that a given information processing module uses, rather than with reducing the uncertainty of the output of the module. The goal of a strongly coupled fusion method is to alter the constraints used by a module based on information from an independent module in such a way as to make the constraints more valid in a given environment. Strongly coupled data fusion involves alteration of either the image formation model, prior model, or system model.

Recurrent strong coupling occurs when the output of a given

Recurrent strong coupling occurs when the output of a given module is fed back, through some path, to affect its own prior constraints. Common examples of recurrent strong coupling are to be found in some Kalman filter based fusional methods, and in coupled Markov Random Field based methods. In the Kalman filter methods, the output of a module is used to change its own prior model, while in the coupled MRF approaches the output of one module affects the prior model of another module, whose output in turn affects the prior model of the original module.

Early approaches to computer vision were almost entirely concerned with the analysis of static images (i.e. snapshots). It is clear, however, that the information processing systems of organisms or machines that move about and interact with their environment need to be concerned with the temporal properties of the structures in the environment. We must therefore be aware of the temporal aspects of the environment when developing data fusion algorithms. In the Bayesian formulation of information processing the temporal aspect of the environment is handled by specifying temporal constraints on the configuration of the world. We described a way to embed a generalized bandlimit constraint on world parameters, and how this leads to an optimal Bayesian estimate of the world parameter. In general, temporal constraints on world parameters will be more complex than the generalized bandlimit constraints we talk about, and computer vision researchers will have to devote no small amount of effort in determining suitable forms for these.

Faced with the prospect of not being able to specify the right models or constraints for a given environment, we would like a way to "learn" these constraints. We presented a methodology for doing this based on the use of consistency measures, or meta-constraints, obtained from the outputs of independent sensory modules, to adapt the prior constraints used by the modules in a way which maximizes the consistency measures. This results in a strongly coupled fusion algorithm, whereby the constraints of a module are modified based on measures of the consistency between the outputs of multiple modules.

The material presented in this text sketches out a number of reasons for why data fusion methods are required. From these reasons we can begin to develop a general theory of data fusion. This theory is based on the idea that the fundamental role of data fusion processes is not to reduce uncertainty in parameter estimates, but rather to ensure that the constraints that are used in the process of inverting the world-image mapping are sufficient and, most importantly, *valid*. The theory behind data fusion should, in this view, be mainly concerned with issues related to these constraints. These issues include that of properly modeling the world or environment, both spatially and temporally, embedding of the constraints in algorithmically efficient ways, determination of effective ways of computing the validity of constraints, and techniques for modifying constraints.

It is the authors hope that this text will cause some researchers to begin looking at the data fusion problem in terms of the constraints needed to solve sensory information processing tasks, rather than just as a method for reducing the effects of sensor noise.

Bibliography

[1] Aloimonos, J., Basu, A., "Combining information in low-level vision.", CAR-TR-336 and CS-TR-1947, University of Maryland Center for Automation Research Technical Report, November 1987

[2] Aloimonos, J., Weiss, I. and Bandyopadhyay, A. "Active Vision", in *Proceedings of the First International Conference on Computer Vision.* June 1987, pp 35-54

[3] Aloimonos, J. and Schulman, D., **Integration of visual modules: An extension of the Marr paradigm,**also in *Proceedings of the 1989 Darpa Image Understanding Workshop,* pp 507-551

[4] Ayache, N., and Faugeras, O. "Building, registrating and fusing noisy visual maps.", in *Proceedings of the First International Conference on Computer Vision,* London, 1987, pp 73-82

[5] Bajcsy, R., "Integrating vision and touch for grasping of an object", GRASP LAB Technical Report 29, Dept. of Computer and Information Science, University of Pennsylvania, 1984.

[6] Bajcsy, R., "Perception with feedback", *Proceedings of the 1988 Darpa Image Understanding Workshop,* pp 279-288

[7] Barnard, S. *Proc. Image Understanding Workshop,* Los Angeles, 1986.

[8] Barnard, S., "A stochastic approach to stereo vision", *Proc. 5th National Conference on Artificial Intelligence*,Philadelphia, pp 676-680, 1986

[9] Barnard, S. and Fischler, M.A, "Computational Stereo". *Computing Surveys*, **14**, No. 4, 1982.

[10] Bayes, T., "An essay towards solving a problem in the doctrine of chances", *Philosophical Transactions of the Royal Society*, Vol. 53, pp 370-418, 1783

[11] Berger, J.O., **Statistical Decision Theory and Bayesian Analysis**, Springer-Verlag, NY, 1985

[12] Bestavros, A., Clark, J.J., and Ferrier, N., "Management of sensori-motor activity in mobile robots", *Proceedings of the 1990 IEEE Conference on Robotics and Automation*, Cinncinati, 1990

[13] Binford, T.O., Baker, "Depth from edges and intensity based stereo", in *Proceedings of the International Joint Conference on Artificial Intelligence*, Vancouver, B.C. pp 631-636. 1981.

[14] Blake, A, "The least-disturbance principle and weak constraints", in *Pattern Recognition Letters*, Vol 1, Nos 5,6. 1983.

[15] Blake, A. "Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 1, pp 2-12, 1989.

[16] Blake, A., and Zisserman, A., '**Visual Reconstruction**, MIT Press, Cambridge, MA, 1987

[17] Blake, I.F., and Lindsey, W.C., "Level-crossing problems for random processes", *IEEE Transactions on Information Theory*, Vol. 19, No. 3, pp 295-315, 1973

[18] Bolle, R.M. and Cooper, D.B., "On optimally combining pieces of information, with application to estimating 3-D complex object position from range data", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8, No. 5, pp 619-638, 1986

[19] Bove, V.M., "Discrete fourier transform based depth from focus", *Optical Society of America Topical Meeting on Image Understanding and Machine Vision*, Cape Cod, MA, June 1989

[20] Brady, J.M. and Yuille, A.L., "An extremum principle for shape from contour", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, pp 288-301. 1984.

[21] Brandt, A., "Multilevel adaptive solutions to boundary value problems", *Mathematical Computing*, Vol. 31, pp 333-390, 1977

[22] Brelstaff, G., and Blake, A., "Detecting specular reflections using Lambertian constraints", *Proceedings of the 2nd International Conference on Computer Vision*, Tampa, 1988, pp 297-302

[23] Brooks, M.J., and Horn, B.K.P., "Shape and source from shading", MIT AI Lab Memo TR-820, 1985

[24] Bülthoff, H. and Fahle, M. "Disparity Gradients and Depth Scaling", MIT AI Lab Memo TR-1175, 1989.

[25] Bülthoff, H. and Mallot H.P., "Interaction of different modules in depth perception", in *Proceedings of the First International Conference on Computer Vision*, London, 1987, pp 295-305

[26] Bülthoff, H. and Mallot, H-P. "Integration of depth modules: stereo and shading". *J. Opt. Soc. Am.*, **5**, 1749-1758, 1988.

[27] Burt, P. and Julesz, B. "A disparity gradient limit for binocular fusion", *Science* **208**, 615-617, 1980.

[28] Canny, J.F., " Finding Edges and Lines in Images", MIT AI Lab Memo, TR-720, 1983

[29] Carnap, R., **The Nature and Application of Inductive Logic**, University of Chicago Press, Chicago, 1951

[30] Chou, P.B., and Brown, C.M., "Multimodal reconstruction and segmentation with Markov random fields and HCF optimization", in *Proceedings of the 1988 Darpa Image Understanding Workshop*, pp 214-221

[31] Clark, J.J., **Multiresolution Stereo Vision with Application to the Automated Measurement of Logs,** Ph.D. Thesis, The University of British Columbia, Sept., 1985

[32] Clark, J.J., "Sampling and reconstruction of non-bandlimited signals", in *Proceedings of Visual Communications and Image Processing IV, SPIE, Philadelphia,* 1989

[33] Clark, J.J., Palmer, M.R., and Lawrence, P.D., "A transformation method for the reconstruction of functions from nonuniformly spaced samples.", *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 33, No. 4, pp 1151-1165, October, 1985.

[34] Clark, J.J., and Yuille, A.L., "Shape from shading via the fusion of specular and Lambertian image components", *Proceedings of the 10th International Conference on Pattern Recognition,* Atlantic City, 1990

[35] Courant, and Hilbert, D., **Methods of Mathematical Physics,** John Wiley, New York, 1953

[36] Definetti, B., "Recent suggestions for the reconciliations of theories of probability", *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* pp 217-26, 1951

[37] Dev, P., "Perception of depth surfaces in random-dot stereograms: A neural model", *International Journal of Man-Machine Studies,* Vol. 7, pp 511-528, 1975

[38] Douglas, R.H., Collett, T., and Wagner, H.J., "Accomodation in *Anuran Amphibia* and its role in depth vision", in *Journal of Comparative Physiology A,* 1986

[39] Drumheller, M. "Mobile robot localization using sonar", MIT A.I. Lab Memo TR-826. 1985.

[40] Duchon, J. **Lecture Notes in Mathematics.** 571. (Eds Schempp, W. and Zeller, K.), 85-100, Springer-Verlag, Berlin, 1979

[41] Durbin, R., Szeliski, R. and Yuille, A.L. "The elastic net and the travelling salesman problem", Harvard Robotics Laboratory Technical Report. No. 89-3, 1989

[42] Durbin, R. and Willshaw, D. "An analog approach to the travelling salesman problem using an elastic net method". *Nature,* Vol. 326, pp 689-691, 1987

[43] Forsythe, D.A., "A novel approach to colour constancy", *Proceedings of the 2nd IEEE Conference on Computer Vision,* 1988, pp 9-18

[44] Frankot, R.T., and Chellappa, R., "A method for enforcing integrability in shape from shading algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 10, No.4, pp 439-451, 1988

[45] Gamble, E. and Poggio, T., "Visual integration and detection of discontinuities: The key role of intensity edges", MIT AI Lab Memo No. TR-970, October 1987

[46] Geiger, D. and Girosi, F., "Parallel and deterministic algorithms from MRFs :integration and surface reconstruction", MIT AI Lab Memo No. TR-1114, June 1989.

[47] Geiger, D. and Poggio, T. "Optimal scale for edge detection", *International Joint Conference on Artificial Intelligence,* Milan. 1987.

[48] Geiger, D. and Yuille, A., "Stereopsis and eye movement", *Proceedings of the First International Conference on Computer Vision,* London, 1987, pp 306-314

[49] Geiger, D. and Yuille, A., "A common framework for image segmentation", Harvard Robotics Laboratory Technical Report. No. 89-7, 1989, also in ICPR '90, Atlantic City, N.J.

[50] Gelb, A. (ed.), **Applied Optimal Estimation,** MIT Press, Cambridge, MA, 1974

[51] Geman, S. and Geman, D. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-6, pp 721-741, 1984

[52] Gennert, M., **A Computational Framework for Understanding Problems in Stereo Vision,** PhD Thesis, Massachusetts Institute of Technology, AI Laboratory, 1987

[53] Gershon, R., **The Use of Color in Computation Vision,** PhD Thesis, Department of Computer Science, University of Toronto, 1987

[54] Gibson, J.J., **Reasons for Realism: Selected Essays of James J. Gibson,** E. Reed and R. Jones (eds.), L. Erlbaum, Hillsdale, N.J., 1982

[55] Glauber, R., "Time dependent statistics of the Ising model", *Journal of Mathematical Physics,* Vol. 4, p294, 1963

[56] Good, I.J., in **The Foundations of Statistical Inference,** Godambe, V.P., and Sprott, D.A., eds., Holt, Rhinehart and Winston, Toronto, 1973

[57] Grimson, W.E.L., **From Images to Surfaces: A Computational Study of the Human Early Vision System,** MIT Press, Cambridge MA, 1981

[58] Guzman, A., "Decomposition of a visual scene into three dimensional bodies", *AFIPS Conference Proceedings,* Vol. 33, pp 291-304, 1968

[59] Henderson, T., Weitz, E., Hansens, C., and Mitiche, A., "Multisensor knowledge systems: Interpreting 3D structure", *International Journal of Robotics Research,* Vol. 7., No. 6, pp 114-137, 1988

[60] Hildreth, E.C., "Computation of the velocity field.", *Proceedings of the Royal Society of London, B,* Vol. 221, pp 189-220, 1984

[61] Hopfield, J.J. and Tank, D.W. "Neural computation of decisions in optimization problems", *Biological Cybernetics,* Vol. 52, 141-152, 1985.

[62] Horn, B.K.P. "Obtaining shape from shading information", Chap. 4, in **The Psychology of Computer Vision,** P.H. Winston, ed., McGraw-Hill Book Company. New York. pp 115-155. 1975.

[63] Horn, B.K.P., **Robot Vision,** MIT Press, Cambridge MA, 1986

[64] Horn, B.K.P., and Brooks, M.J., "The variational approach to shape from shading.", MIT AI Lab memo No. TR-813, March 1985

[65] Horn, B.K.P., and Schunk, B., "Determining optical flow", *Artificial Intelligence,* Vol. 17, No. 1-3, p185, 1981

[66] Horn, B.K.P. and Sjoberg, R.W., "Calculating the reflectance map", *Applied Optics,* Vol. 18, No. 11, pp 1770-1779

[67] **House, D.H. Neural models of depth perception in frogs and toads.** Ph.D. Thesis, The University of Massachusetts at Amherst, Sept. 1984.

[68] Hutchinson, J., Koch, C., Luo, J. and Mead, C. "Computing motion using analog and binary resistive networks", Caltech Preprint. 1987.

[69] Hwang, Ten-Lee., Clark, J.J. and Yuille, A.L. "A depth recovery algorithm using defocus information.", Harvard Robotics Laboratory Technical Report, No. 89-2, 1989.

[70] Hwang, Ten-Lee., and Clark, J.J., "A spatio-temporal generalization of Canny's edge detector", *Proceedings of the 10th International Conference on Pattern Recognition,* Atlantic City, 1990

[71] Hwang, Ten-Lee., and Clark, J.J., "On local detection of moving edges", *Proceedings of the 10th International Conference on Pattern Recognition,* Atlantic City, 1990

[72] Ikeuchi, K., and Horn, B.K.P., "Numerical shape from shading and occluding boundaries", *Artificial Intelligence,* Vol. 17, No. 1-3, pp 141-184, 1981

[73] Ikeuchi, K., "Shape from regular patterns", *Artificial Intelligence,* Vol. 22, No. 1, pp 49-75, 1984

[74] Jepson, A.D. and Jenkin, M.R.M. "The fast computation of disparity from phase differences", *Proceedings of the 1989 Computer Vision and Pattern Recognition,* San Diego, pp 398-403, 1989

[75] Jerri, A.J., "The Shannon sampling theorem - Its various extensions and applications: A tutorial review", *Proceedings of the IEEE,* Vol. 65, No. 11, pp 1565-1596, 1977

[76] Kahn, P., "Integrating moving edge information along a 2D trajectory in densely sampled imagery.", *Proceedings of the 1988 Computer Vision and Pattern Recognition Conference,* Ann Arbor, pp 702-709, 1988

[77] Kant, I., **Critique of Pure Reason,** St. Martin's Press, NY, 1929

[78] Keeler, K., "Universal coding and symbolic inference with order information", Harvard Robotics Laboratory Technical Report, No. 90-1, 1990

[79] Kirkpatrick, S., Gelatt, C.D. Jr. and Vecchi, M.P. "Optimization by simulated annealing", *Science,* **220**, pp 671-680, 1983

[80] Klinker, G.J., Shafer, S.A., and Kanade, T., "Color image analysis with an intrinsic reflection model", in *Proceedings of the 2nd International Conference on Computer Vision,* Tampa, Florida, pp 292-296, 1988

[81] Koch, C., Marroquin, J. and Yuille, A.L. "Analog "neuronal" networks in early vision", *Proceedings of the National Academy of Science, U.S.A.,* Vol. 83, pp 4263-4267, 1986

[82] Kramer, H.P., "A generalized sampling theorem", *Journal of Mathematical Physics,* Vol. 38, pp 68-72, 1959

[83] Krotkov, E., and Kories, R., "Adaptive control of cooperating sensors: Focus and stereo ranging with an agile camera system", in the *Proceedings of the 1988 IEEE Conference on Robotics and Automation,* Philadelphia, pp 548-553, 1988

[84] Land, E., and McCann, J.J., "Lightness and retinex theory", *Journal of the Optical Society of America,* Vol. 51, No. 1, pp 1-11, 1975

[85] Leclerc, Y., "Constructing simple stable descriptions for image partitioning", *International Journal of Computer Vision,* Vol. 3, No. 1, pp 75-102, 1989

[86] Lee, H.-C., "Method for computing the scene-illuminant chromaticity from specular highlights", *Journal of the Optical Society of America A,* Vol. 3, No. 10, pp 1694-1699, 1986

[87] Lee, S.J., Haralick, R.M. and Shapiro, L.G. "Morphological edge detection", Technical Report, Machine Vision International, Ann Arbor, Michigan 48104, 1986

[88] Levinson, N., **Gap and Density Theorems,** Colloquim Publication 26, American Mathematical Society, New York, 1940

[89] Locke, J., "An essay concerning human understanding", in **Locke's Essays,** James Kay, Jun. and Brother, Philadelphia, 1927

[90] Longuet-Higgins, H.C., "The role of the vertical dimension in stereoscopic vision", *Perception,* Vol. 11, pp 377-386, 1982

[91] Longuet-Higgins, M.S., "The distribution of intervals between zeroes of a stationary random function", *Proceedings of the Royal Society of London, A,* Vol. 254, pp 557-599, 1962

[92] Luenberger, D.G., **Introduction to Linear and Nonlinear Programming,** Addison-Wesley, Reading, MA, 1973

[93] Lumsdaine, A., Wyatt, J., and Elfadel, I., "Nonlinear analog networks for image smoothing and segmentation", *International Symposium on Circuits and Systems,*1990

[94] Lunscher, W.H.H.J., **A digital image preprocessor for optical character recognition,** MASc Thesis, Department of Electrical Engineering, University of British Columbia, 1983

[95] Maloney, L.T. and Wandell, B.A., "A computational model of color constancy", *Journal of the Optical Society of America,* pp 29-33, 1986

[96] Maragos, P. **A unified theory of tranlation invariant systems with applications to morphological analysis and coding of images.** Ph.D. Thesis. Georgia Institute of Technology. July, 1985.

[97] Marr, D., **Vision,** Freeman, San Francisco, 1982

[98] D. Marr and E.C. Hildreth,"Theory of edge detection", *Proceedings of the Royal Society of London, series B,* Vol. 207, pp 187-217, 1980

[99] Marr, D. and Poggio, T., "Cooperative computation of stereo disparity", *Science,* No. 194, pp 283-287

[100] Marr, D., and Poggio, T., "A computational theory of human stereo vision", *Proceedings of the Royal Society of London B,* Vol. 204, pp 301-328

[101] Marroquin, J., **Probabilistic solutions of inverse problems.** Ph.D. Thesis. MIT. Sept, 1985.

[102] Marroquin, J. "Deterministic Bayesian estimation of Markovian random fields with applications to computational vision", in the *Proceedings of the First International Conference on Computer Vision*, London, 1987, pp 597-601

[103] Marroquin, J., Mitter, S., and Poggio, T., "Probabilistic solutions of ill-posed problems in computational vision", *Journal of the American Statistical Association,* Vol. 82, No. 397, March 1987, pp 76-89

[104] Maund, C., **Hume's Theory of Knowledge,** Russell and Russell, N.Y., 1972

[105] Matthies, L., Kanade, T., and Szeliski, R., "Kalman filter-based algorithms for estimating depth from image sequences", *International Journal of Computer Vision,* Vol. 3, pp 209-236, 1989

[106] Matthies, L., Szeliski, R., and Kanade, T., "Incremental estimation of dense depth maps from image sequences", *IEEE Conference on Computer Vision and Pattern Recognition,* pp 366-374, 1988

[107] Metropolis, N. Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. "Equation of state calculations by fast computing machines", *Journal of Physical Chemistry,* Vol. 21, pp 1087-1091, 1953.

[108] Michaels, C., and Carello, C., **Direct Perception,** Prentice-Hall, Englewood Cliffs, N.J., 1981

[109] Middleton, D., **An Introduction to Statistical Communication Theory,** Peninsula Publishing, Los Altos CA, 1987

[110] Miller, K.S., **Multidimensional Gaussian Distributions,** John Wiley and Sons, New York, 1964

[111] Mitchison, G.M. "Planarity and segmentation in stereoscopic matching", *Perception,* Vol. 17, pp 753-782, 1988.

[112] Mitchison, G.M. and McKee, S. "The resolution of ambiguous stereoscopic matches by interpolation", *Vision Research,* Vol 27, no 2. pp 285-294, 1987

[113] Mumford, D. and Shah, J., "Boundary detection by minimizing functions", in the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* San Fransisco, 1985

[114] Nandakumar, N. and Aggarwal, J.K., "Multisensor fusion for scene perception - Integrating thermal and visual imagery.", Technical Report TR-87-9-41, Computer and Vision Research Center, University of Texas at Austin, August 1987

[115] Nandakumar, N. and Aggarwal, J.K., "Multisensor integration - Experiments in integrating thermal and visual sensors.", in the *Proceedings of the 1st IEEE Conference on Computer Vision,* London, pp 83-92, 1987

[116] Nowlin, W.C., "Bayes theorem and unimodality of products of pdf's with an application to tactile sensing", Harvard Robotics Laboratory Technical Report, No. 89-15, 1989

[117] Ohta, Y., Watanabe, M., and Ikeda, K. "Improving depth map by right-angled trinocular stereo", *Proceedings of the IEEE International Conference on Pattern Recognition, Paris,* pp 519-521, 1986

[118] Paley, R.E.A.C. and Wiener, N., **Fourier Transforms in the Complex Domain,** Colloquim Publication 19, American Mathematical Society, New York, 1934

[119] Parisi, G. **Statistical Field Theory**, Addison-Wesley, Reading, Mass. 1988

[120] Pentland, A., "A new sense for depth of field", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 7, No. 2, 1985

[121] Pollard, S.B., Mayhew, J.E.W. and Frisby, J.P. "Disparity Gradients and Stereo Correspondences", *Perception,* 1987

[122] Poggio, T., and staff, "MIT progress in understanding images.", in the *Proceedings of the Image Understanding Workshop,* Los Angeles, 1987

[123] Poggio, T. Gamble, E.B., and Little, J.J., "Parallel integration of visual modules", *Science,* Vol. 242, pp 436-440, October 1988

[124] Poggio, T. and Koch, C., "An analog model of computation for the ill-posed problems of early vision.", MIT AI memo No. TR-783, May 1984

[125] Poggio, T., and Torre, V., "Ill posed problems and regularization analysis in early vision.", MIT AI memo No. TR-773, 1984

[126] Poggio, T., Voorhess, H., and Yuille, A., "A regularized solution to edge detection.", MIT AI Lab memo No. TR-833, May 1985

[127] Porrill, J. Private Communication.

[128] Prasad, K.V., Mammone, R.J., and Yogeshwara, J., "3-D image restoration using constrained optimization techniques", *Optical Engineering,* to be published.

[129] Prazdny, K., "Detection of binocular disparities.", *Biological Cybernetics,* Vol. 52, pp 93-99, 1985

[130] Reichenbach, H., **The Theory of Probability,** University of California Press, Berkeley, 1949

[131] Rice, S.O., "Mathematical analysis of random noise", *Bell System Technical Journal,* Vol. 24, pp 46-156, 1945

[132] Richardson, J.M., and Marsh, K.A., "Fusion of multisensor data", *International Journal of Robotics Research,* Vol. 7., No. 6, pp 78-96, 1988

[133] Risannen, J., "A universal prior for integers and estimation by minimum description length", *Annals of Statistics,* Vol. 11, No. 2, pp 416-431, 1983

[134] Roberts, L.G., "Machine perception of three-dimensional solids", in **Optical and Electro-optical Information Processing,** J.P. Tippett et al (Eds.), MIT Press, Cambridge, MA, 1965

[135] Rosenfeld, A., Hummel, R., and Zucker, S., "Scene labelling by relaxation operations", *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 6, p 420, 1976

[136] Sanger, T. "Stereo disparity computation using Gabor filters," *Biological Cybernetics,* **59**, 405-418, 1988

[137] Savage, L.J., "The subjective basis of statistical practice", Technical Report, Department of Statistics, University of Michigan, 1961.

[138] Savage, L.J., **The Foundations of Statistics**, 2nd Edition, Dover, New York, 1972

[139] Shafer, S., "Using colour to separate reflection components", Technical Report TR-136, Computer Science Dept, University of Rochester.

[140] Shannon, C.E., "Communications in the presence of noise", *Proceedings of the IRE*, Vol. 37, pp 10-21, 1949

[141] Shlomot, E., and Zeevi, Y.Y., "Nonuniform sampling and representation of images which are not bandlimited", EE Pub. No. 742, Department of Electrical Engineering, Technion Israel Institute of Technology, 1990

[142] Singh, A., "Image-flow estimation: An information fusion approach", in the *Proceedings of the 1989 Darpa Image Understanding Workshop*, pp 983-991, 1989

[143] Stansfield, S.A., "Visually aided tactile exploration", in the *Proceedings of the 1987 IEEE Robotics and Automation Conference*, Raleigh NC, pp 1487-1492, 1987

[144] Stevens, K., **Surface Perception From Local Analysis of Texture and Contour**, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979

[145] Subbarao, M., "Parallel depth recovery by changing camera parameters", *Second IEEE International Conference on Computer Vision*, pp 149-155, 1988

[146] Szeliski, R., "Estimating motion from sparse range data without correspondence", *2nd International Conference on Computer Vision*, pp 207-216, 1988

[147] Szeliski, R., **Bayesian Modelling of Uncertainty in Low-level Vision**, Kluwer, Boston, 1989

[148] Terzopoulos, D., "The role of constraints and discontinuities in visible surface reconstruction.", in the *Proceedings of the International Joint Conference on Artificial Intelligence*, Karlsruhe, 1983

[149] Tikhonov, A.N., and Arsenin, V.Y., **Solution of Ill-posed Problems**, Winston, Washington D.C., 1977

[150] Ullman, S., **The interpretation of visual motion**, MIT Press. Cambridge, Mass. 1979

[151] Ullman, S., "Maximizing rigidity: the incremental recovery of 3-D structure from rigid and rubbery motion", *Perception*, Vol. 13, pp 255-274, 1984

[152] von Mises, R., **Probability, Statistics and Truth**, MacMillan, New York, 1957

[153] Wahba, G., "Practical approximate solutions to linear operator equations when the data are noisy", *SIAM Journal on Numerical Analysis*, Vol. 14, No. 4, 1977

[154] Wald, A., **Statistical Decision Functions**, Wiley, New York, 1950

[155] Warrington, E.K., and Taylor, A.M., "The contribution of the right parietal lobe to object recognition", *Cortex*, Vol. 9, pp 152-164, 1973

[156] Wasserstrom, E. "Numerical solutions by the continuation method", *SIAM Review*, Vol. 15, 89-119, 1973

[157] Waxman, A., and Wohn, K., "Contour evolution, neighborhood deformation and global image flow: planar surfaces in motion.", *International Journal of Robotics Research*, Vol. 4, No. 3, pp 95-108

[158] Weber, J.D., **Historical Aspects of the Bayesian Controversy**, The University of Arizona, Tucson Arizona, 1973

[159] Witkin, A. "Recovering surface shape and orientation from texture", *Artificial Intelligence*, Vol. 17, pp 17-47, 1981

[160] Wolff, L., "Using polarization to separate reflectance components", *Proceedings of the 1989 IEEE Conference on Computer Vision and Pattern Recognition,* San Diego, pp 361-369, 1989

[161] Wong, E., "Two dimensional random fields and representation of images", *SIAM Journal of Applied Math,* Vol. 16, No. 4, pp 756-770, 1968

[162] Woodham, R.J., "Photometric stereo: A reflectance map technique for determining surface orientation from image intensity", *Image Understanding Systems and Industrial Applications, Proc. SPIE 155,* 1978

[163] Wu, J., **Motion Estimation from Image Sequences,** PhD thesis, Division of Applied Sciences, Harvard University, 1987

[164] Yuille, A.L., "A method for computing spectral reflectance", *Biological Cybernetics,* Vol. 56, pp 195-201, 1987

[165] Yuille, A.L., and Grzywacz, N.M., "A computational theory for the perception of coherent visual motion", *Nature,* May 1988

[166] Yuille, A.L., "Deformable templates, mean field theory and matching", Harvard Robotics Laboratory Technical Report 89-12, 1989

[167] Yuille, A.L., "Energy functions for early vision and analog networks", *Biological Cybernetics,* Vol. 61, pp 115-123, 1989

[168] Yuille, A.L., Cohen, D., and Hallinan, P., "Facial feature extraction by deformable templates", Harvard Robotics Laboratory Technical Report, No. 88-2, 1988

[169] Yuille, A.L., Geiger, D., and Bülthoff, H.H., "Stereo integration, mean field theory and psychophysics",*First European Conference on Computer Vision,* Antibes, France 1990

[170] Yuille, A.L. and Gennert, M., "A theory for coherent stereo", preprint, 1988

[171] Yuille, A.L., Yang, T., and Geiger, D., "Towards a theory of transparency" in preparation.

Index

active vision, 8, 75, 88, 90, 92, 209
affordances, 2, 11, 33
Aggarwal, J., 90
aliasing error, 204
Aloimonos, J., 88
average risk, 29–30
Ayache, N., 95

Barnard, S., 107–108, 114
basis functions, 57, 59, 65
Bayes rule, 18
Bayesian
  controversy, 30
  interpretation, 17
Blake, A., 45, 55
blocks world, 12
Brandt, A., 40
Brooks, M.J., 158
Brown, C., 102
Bulthoff, H., 105–107, 113

Canny, J., 199
Chou, P., 102
classification, 30
Collett, T., 98, 142
color, 148
color constancy, 57
constraint
  a priori, 24, 32–36, 44, 57, 63–65
  adaption, 25, 64, 79
  artificial, 7, 32, 71
  bandlimit, 202
  compatibility, 109
  continuity, 109
  convection, 98
  determination, 9
  domain specific, 23
  elastic membrane, 41, 56
  embedding, 9, 71
  epipolar, 111
  generalized bandlimit, 202
  global, 54
  integrability, 159
  interaction, 56
  Lambertian, 7
  matching, 48, 51, 107, 111
  minimum area, 174, 177
  multisubjective, 36, 67, 173
  natural, 7, 12, 32, 35, 71, 111
  neighborhood, 98
  parametric, 57, 65–66
  physical, 7, 12, 20, 32, 34, 71, 111
  prior, 6, 76
  rigidity, 5, 7, 208

smoothness, 5, 7, 17, 39, 42, 45, 47, 55–57, 110–111, 160, 162, 174
temporal, 181, 212
temporal coherence, 182
universal, 13
weak, 55
world-image mapping, 5
correspondence problem, 106
cover principle, 47–48
cross-validation, 41, 212

data fusion, 13, 67, 71
data fusion
algebraic, 75, 90, 99, 150
class I weakly coupled, 73
class II weakly coupled, 75, 90, 150
class III weakly coupled, 76, 99, 158
constraint, 36
recurrent, 79, 101, 177
strongly coupled, 25, 36, 45, 48, 64, 67, 72, 78, 97, 101, 113, 168
weakly coupled, 63, 67, 72
decision theory, 29, 175
deformable template, 59
depth from defocus, 98, 213
detection, 29
Dev, P., 142

ecological optics, 8
energy
effective, 55, 107
minimization, 18, 39, 44, 47, 51, 59, 80, 85, 93, 111, 115, 158
estimator
Bayesian, 29
Conditional Mean, 26, 29, 50
Maximum a Posteriori, 25, 28, 50
Maximum Likelihood, 25, 29
Minimum Mean Square Error, 25
Minimum Variance, 25, 50, 116
Euler-Lagrange equations, 41–42, 48, 55, 161–162, 169
extraction, 29

face recognition, 59
Fahle, M., 106
Faugeras, O., 95
figural continuity, 111
Forsythe, D., 59
Frisby, J., 108

Gamble, E., 102
Geiger, D., 50, 54, 97, 105, 138
Geman, D., 45–46, 54, 82
Geman, S., 45–46, 54, 82
Gennert, M., 107–108, 114
Gibb's distribution, 43–45, 50, 54, 57, 84, 114, 160
Gibson, J.J., 2, 8, 11, 33
Girosi, F., 50, 54
Glauber, R., 53
Good, I.J., 33
gradient descent, 40, 63
Green's function, 42–43, 48, 56

Grimson, W.E.L., 5, 40, 106
Grzywacz, N.M., 42, 107

Hildreth, E., 40
Hopfield, J., 107
Horn, B.K.P., 40, 98, 158
House, D.H., 98, 138, 142, 144
Hume, D., 10
Hummel, R., 39
Hwang, T.L., 146, 199

Ikeuchi, K., 40
ill-posed problems, 6, 14, 23, 33, 88
instability, 155

Jenkin, M., 108
Jepson, A., 108

Kalman filter, 27, 92, 96–97, 182
Kalman filter
extended, 95
Kant, I., 9
Karhunen-Loeve matrix, 65
Keeler, K., 66
Kories, R., 214
Kramer, H.P., 203
Krotkov, E., 214

Leclerc, Y., 66
Levinson, N., 203
likelihood ratio, 30
line process, 45–46, 54–56, 102, 107, 112, 174
Little, J., 102
Longuet-Higgins, H.C., 139
Lumsdaine, A., 54, 107
Lyapunov function, 61

Mallot, H.P., 107, 113
Markov Random Field, 18, 44, 57, 105
Markov Random Field
coupled, 45–46, 102
Marr, D., 2, 35, 108, 112, 142
Marroquin, J., 50
matching, 46–47, 106–107
Matthies, L., 24–25, 92
Mayhew, J., 108
mean field, 50, 115
mean field approximation, 51–52
minimal mapping, 47–49
minimal surface, 165
minimum description length, 18, 64
minimum length coding, 11
model
dichromatic, 148
image formation, 12, 19, 27, 32, 34, 64, 71, 79, 101, 174, 182, 213
Lambertian, 21
prior, 12, 23–25, 27, 64, 71, 79, 93, 102, 115, 201
probabilistic, 12
sensor, 21
system, 24–26
temporal, 182
uncertainty, 21
motion
analysis, 43, 98
correspondence, 39

depth from, 92
from depth, 93
long range, 46
long-range, 107

Nandhakumar, N., 90

Paley, R.E.A.C., 203
partition function, 50–52, 54
photometric stereo, 89
Poggio, T., 40, 84, 102, 108, 142
polarization, 148
Pollard, S., 108
Prazdny, K., 108
probability
objective, 31, 33
subjective, 31–33, 35

reflectance, 20
reflectance
Lambertian, 147, 149
specular, 147, 149
reflectance map, 58, 89
regularization, 14, 18, 39–40, 46
Risannen, J., 11, 64
Rosenfeld, A., 39

Sanger, T., 108
Savage, L., 31
Schulman, D., 88
Schunk, B., 40, 98
segmentation, 30, 46, 54, 66, 81–82, 90, 174
Shannon, C., 203, 205
shape from shading, 5, 20, 40, 81, 147
shape from texture, 40
simulated annealing, 46
Singh, A., 98
stereo vision, 78, 97, 105
structure-from-motion, 212
Szeliski, R., 87, 93

Tank, D., 107
temporal consistency, 64
Terzopoulos, D., 40
Torre, V., 40
transparency, 66
traveling salesman problem, 107

Ullman, S., 39, 46
uncertainty, 151, 205

Weber, J.D., 30–31
Wiener, N., 203
winner-take-all, 51
Wolff, L., 148
Woodham, R.J., 89
world-image mapping, 3, 8, 33, 39, 181

Zucker, S., 39