

To the Graduate Council:

I am submitting herewith a dissertation written by Scott M. Thayer entitled "Task-Domain Vision: Theory and Architecture." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Electrical Engineering.

m. A. Abidi

Dr. Mongi A. Abidi, Major Professor

We have read this dissertation
and recommend its acceptance:

Donald W. Boulden

Michael G. Thomas

Marshall Pace

Accepted for the Council:

LeW Minkel

Associate Vice Chancellor and
Dean of The Graduate School

TASK-DOMAIN VISION: THEORY AND ARCHITECTURE

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Scott M. Thayer

May 1998

Abstract

Active vision, as a discipline, is relatively new in the short history of machine vision. It has, however, left an indelible mark in the minds of vision researchers with respect to the design of visual systems. It is the purpose of this research to explore the design of an active component in a purposive, or task-domain, framework and formulate a general theory of task-oriented design for a class of imaging devices. In particular, this theory will address three general areas that are problematic in the design of general vision modules. These are (1) *Sensor Geometry, Calibration, and Representation*. It is shown how application of task-domain constraints can reduce the complexity of these problems and in certain cases present closed-form solutions to optimal design and configuration problems.

In addition, a systemic view of algorithm design is presented. Using this methodology, a Bayesian network of algorithms is used to formulate possible solutions to a particular problem. Using purposive constraints, the most favorable solutions are propagated and the undesirable solutions are discarded. This architecture offers significant advantages over the traditional block-level system design in terms of accuracy and robustness.

Finally, a design philosophy of purposive vision is presented. This methodology is employed in the design of three active visual modules. The three applications are: (1) A parallel-axis stereo system for small object localization; (2) A convergent-axis stereo system with efficient dynamic reconfiguration capabilities; and (3) an active Bayesian network for synthesis of cranial structures in positron emission tomography. The versatility of this theory is evident in its range of applicability to vision problems.

TABLE OF CONTENTS

CHAPTER	PAGE
1. Introduction	1
1.1 Research Domain	2
1.2 Problem Domain	4
1.2.1 Sensor Geometry	6
1.2.2 Calibration	8
1.2.3 Representation	9
1.3 Overview of Research Contributions	10
1.4 Organization	11
2. A Task-Oriented Design Methodology	13
2.1 Pruning the Possibility Tree	14
2.2 MODEL	15
2.3 TRAIN	16
2.4 FOCUS	18
2.5 FIRE	19
2.6 Conclusions	19
3. Literature Review	21
3.1 Introduction	21
3.2 The Active Paradigm	23
3.2.1 Concept I: Selective Data Processing	24

CHAPTER	PAGE
3.2.2 Concept II: Exploratory Vision	27
3.2.3 Concept III: Regularization of Vision Problems	29
3.2.4 Concept IV: Task-Driven Theory	29
3.3 Active System Design	30
3.3.1 Characteristic: Foveation	31
3.3.2 Characteristic: Gaze Control and Attention	31
3.3.3 Characteristic: Integration of Task-Domain Constraints	32
3.4 Active Platforms	33
3.4.1 Early Work	34
3.4.2 The Harvard Head	35
3.4.3 The KTH Head	38
3.4.4 TRISH	41
3.4.5 The Aalborg Head	42
3.4.6 The Surrey Attentive Robot Vision System	47
3.4.7 The University of Pennsylvania Head	49
3.4.8 BETH	51
3.4.9 PENNEYES	54
3.5 Discussion	56
4. Optimal Geometries for Parallel-Axis Stereo	58
4.1 Introduction	58
4.2 A Simple Model for Parallel-Axis Stereo Geometry	60
4.3 Isolating Task-Based Constraints	62
4.3.1 A Rectangular Work Space Constraint	65
4.3.2 A Circular Work Space Constraint	68

CHAPTER	PAGE
4.4 The Parallel-Axis Stereo Synthesis (PASS) Algorithm	69
4.4.1 Specification	70
4.4.2 Optimization	72
4.4.3 Translation	72
4.5 A Geometric Description of the Optimality	73
4.6 Static Parallel-Axis Configuration	74
4.7 Conclusions	77
5. Optimal Geomtries for Convergent Stereo	78
5.1 A Model for Convergent-Axis Stereo Geometry	78
5.1.1 Solving for the Stereoscopic Region Vertex Points	81
5.1.2 Derivation of Convergent Stereo Area Equation	81
5.2 Isolating Task-Domain Constraints	83
5.2.1 A Circular Work Space Constraint	84
5.3 Nonlinear Constrained Optimization for Convergent Stereo Configurations	87
5.3.1 Static Baseline Configurations	88
5.3.2 Dynamic Baseline Configurations	89
5.4 The Convergent-Axis Stereo Synthesis (CASS) Algorithm	92
5.4.1 Specification	92
5.4.2 Optimization	93
5.4.3 Translation	95
5.5 A Geometric Description of the Optimality	96
5.6 Conclusions	97
6. Calibration for Integrated Active Stereo	99

CHAPTER	PAGE
6.1 Introduction	99
6.2 Calibration Mathematics	101
6.2.1 Derivation of A_{nci}	102
6.2.2 Selection of Solution Algorithms	103
6.2.3 Selection of Basis Polynomials	104
6.3 Monocular Error Analysis	109
6.3.1 Geometric Interpretation of Monocular Error	110
6.4 Binocular Error Analysis	114
6.4.1 Perturbation Analysis	117
6.5 Conclusions	124
7. Representations for Integrated Active Stereo	126
7.1 Representational Philosophies	127
7.1.1 Challengers with Purpose: The Purposivist	127
7.1.2 PeaceMakers: The Realists	129
7.1.3 What has been Learned	131
7.2 Parametric Elliptical Fourier Representations	132
7.2.1 PEFR's as a Representation for Active Stereo	135
7.2.2 Fundamental PEFR Operators	139
7.3 Conclusions	141
7.3.1 Computational Efficiency	141
7.3.2 Storage Requirements	142
7.3.3 Direct Mapping in Computation Theory	143
8. Architecture for Purposive System Design	145

CHAPTER	PAGE
8.1 Sequential Block Algorithm Design	146
8.1.1 General Algorithms and SBA	149
8.1.2 Looking for a Better Framework	150
8.2 Parallel Block Algorithm Design	151
8.2.1 SBA Networks	151
8.3 Bayesian Algorithmic Network Design	154
8.3.1 A Bayesian Model	155
8.3.2 A Task-Domain Estimator	156
8.3.3 A Simple Example	160
8.4 Conclusions	163
9. Dual Arm Work Module	165
9.1 Implementation of a PASS System	166
9.2 Simulations of the PASS Model	167
9.3 Experimental Validation	171
10. A Convergent-Axis Stereo System	174
10.1 Hardware Design	175
10.2 Computational Architecture	176
10.3 Pipelined Auto-Focus	178
10.3.1 An Optimal Technique	179
10.3.2 A Real-Time Implementation	179
10.3.3 Results and Conclusions	180
10.4 Pipelined Auto Aperture	181
10.4.1 An Optimal Technique	182

CHAPTER	PAGE
10.4.2 A Weakly Calibrated Technique for Auto Aperture	184
10.4.3 Results	185
10.4.4 Conclusions	186
10.5 Dynamic Regional Reconfiguration	187
10.6 Range Estimation	190
10.7 Conclusion	190
11. Automated Synthesis of Structure from Cranial Imagery using BAND	
Networks	192
11.1 Introduction	192
11.2 Volume Segmentation	195
11.3 Iterative Contour Generation Algorithm	196
11.4 Boundary Localization in the Mid Brain:	198
11.5 Boundary Localization in the Superior Brain	201
11.6 Partial Volume Effects	203
11.7 Volume Synthesis	203
11.8 Attenuation Sinogram Generation	205
11.9 Experimental Results	207
12. Discussion and Conclusions	212
BIBLIOGRAPHY	217
APPENDICES	226
A. Future Research Possibilities	227
A.1 A General Framework for Digital Camera System Design	227

CHAPTER	PAGE
A.2 Integration of Representation into the Acquisition	227
A.3 Extension of the Calibration Model	228
A.4 Use of Generalized Representations	228
A.5 Use of Composite Bayes Networks	229
VITA	230

LIST OF TABLES

TABLE		PAGE
3.1	Primary Mechanical Capabilities of the Human Head.	34
3.2	Primary Mechanical Capabilities of the Harvard Head.	38
3.3	Primary Mechanical Capabilities of the TRISH.	42
3.4	Primary Mechanical Capabilities of the Aalborg Head.	44
3.5	Primary Mechanical Capabilities of the Surrey Head.	48
3.6	Primary Mechanical Capabilities of the Pennsylvania Head. . .	50
3.7	Primary Mechanical Capabilities of BETH.	52
4.1	Parallel Axis Geometry with Circular Work Space: The stand-off is held constant, and the normalized focal length is constant as a consequence of the equilateral geometry of the triangle.	75
4.2	Parallel Axis Geometry with Square Work Space: The standoff is held constant, and the normalized focal length is constant as a consequence of the isoceles geometry of the triangle.	75
4.3	Parallel Axis Geometry with Rectangular Work Space I: The standoff is held constant at and the normalized focal length is constant for each α as a consequence of the isoceles geometry of the triangle. . .	76
4.4	Parallel Axis Geometry with Rectangular Work Space II: The standoff is held constant at and the normalized focal length is constant for each α as a consequence of the isoceles geometry of the triangle. . .	76

TABLE	PAGE
6.1 Typical Condition Numbers vs. Interpolation Polynomial Order: Column 1 reports the order of the interpolation polynomial. Column 2 contains condition numbers for the singular value decomposition while column 3 pertains to the Moore-Penrose pseudo inverse.	104
6.2 Synthesis of Significant Polynomials(High Order): Constraints on matrix condition and residual error produce a vector that describes whether a particular polynomial coefficient can be zeroed. The first column indicates from which camera the data was obtained, while the second column indicates the algorithm.	109
6.3 Synthesis of Significant Polynomials (Mid Order).	110
6.4 Synthesis of Significant Polynomials(Low Order).	111
6.5 Effects of Perturbation on Condition: Simulations where run for 1000 and 1,000,000 iterations respectively.	112
6.6 Geometric Error Analysis: Six sets of test data are analyzed with three different norm functions (mm).	113
6.7 Geometric Error Analysis: Derivation of accuracy and precision. (mm)	113
7.1 Estimation of the Input Curve: with varying orders	135
7.2 Operation Cost Comparison: for three different types of template matching schemas	143
7.3 Storage Cost Comparison: with varying resolutions and orders . . .	144
8.1 Sequential Block Algorithm (4 process algorithm): Showing how a sequence of relatively robust processes can lead to systems with low reliability	148

TABLE	PAGE
8.2 General Algorithm (1 process, 4 data classes): An algorithm that is designed for a particular data set may suffer instabilities when presented with dissimilar types of input	150
8.3 BAND Square Wave Estimator (2 process algorithm): Showing how a sequence of relatively simple algorithms can lead to very robust algorithms	164
9.1 Planar Model Coefficients: Least-Squares coefficients for planar model of two sets of range data obtained from the same three-dimensional reference point.	172
10.1 Planar Model Coefficients: Least-Squares coefficients for planar model of two sets of range data obtained from the same three-dimensional reference point.	190

LIST OF FIGURES

FIGURE	PAGE
1.1 “Black Box” Vision: A generic model of imaging where an unstructured scene is imaged with a standard ranging system.	3
1.2 Abstracted Design Space: A single design cannot possibly represent all possible applications; therefore, design parameters must be tuned for each specific application.	5
1.3 Simple Triangulation: Three-dimensional depth is solved by the intersection of two measurement vectors. The accuracy of the resultant depth is a function of this geometry.	7
1.4 Representations of Thirteen: Different representations make certain types of information explicit at the expense of other information.	9
3.1 Passive Sensory/Perception Cycle: The cycle is strictly open-loop. And, the sensation and perception modules have been abstracted into two separable components.	22
3.2 Active Sensory/Perception Cycle: The cycle is strictly closed-loop. And, control of sensation is executed by the perception module to provide intelligence to the sensation process.	23
3.3 Selective Data Processing: Only a portion of this image contains information of interest to the user. (In this case a structured lighting system). Most of the image data can be disregarded. (Courtesy: F. W. Depiero)	26

FIGURE	PAGE
3.4 Exploratory Vision: An information-dense image. Intelligent systems must have the ability explore these types of scenes in an efficient manner.(S. Dali: April in Paris)	28
3.5 The Harvard Robotic Head: A low cost design for gaze and fixation experimentation.	36
3.6 The KTH Robotic Head: A anthropomorphic design to imitate mammalian head movement capabilities.	39
3.7 The Aalborg Head: A head designed with purposive constraints. . . .	46
3.8 The Surrey Attentive Robot Vision System: A head designed with and integrated engineering specification.	49
3.9 Binocular Eyes for Tennessee Head (BETH): A multi-purpose head for use in tracking and robot convoying.	53
3.10 The PENN EYES Robotic Head: A commercially available head design.	55
4.1 Basic Geometry: Stereo vision system.	61
4.2 Geometrical Design Space: Without meaningful constraints, the design space is infinite. (No stable design points can be located)	62
4.3 Geometrical Design Space: The addition of simple physical constraints offers a more manageable design space; however, the search space has been restricted in a trivial context	64
4.4 Geometrical Design Space: Here the design space has been constrained by a task oriented constraint (standoff: equation);however, each point along the curve can be considered optimal	65

FIGURE	PAGE
4.5 Circumscribing an Isoceles Triangle onto a Rectangle: An infinite number of solutions exist. However, Area and Perimeter have a distinct minima.	66
4.6 Circumscription of an Isoceles Triangle onto a Rectangle: The geometrical relationships superimposed with the task-based constraints give a direct solution to the stereo geometry problem.	67
4.7 Circumscription of an Isoceles Triangle onto a Circle: An infinite number of solutions exist. However, area and perimeter have a distinct minima	68
4.8 Circumscription of an Isoceles Triangle onto a Circle: The geometrical relationships superimposed with the task-based constraints give a direct solution to the stereo geometry.	70
4.9 Parallel-Axis Stereo Synthesis (PASS) Algorithm: Systemic constraints are mapped into unique stereo geometries using a minimum area difference solution.	71
4.10 Identification/Localization Tradeoff: As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.	73
5.1 Convergent Stereo Geometry: As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.	79
5.2 Solution for Area of Quadrilateral: As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.	82

FIGURE	PAGE
5.3 Convergent Stereo Geometry: As the area of the polygon increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.	84
5.4 Convergent Stereo Geometry with Unconstrained Baseline Translation: As the area of the polygon increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.	91
5.5 Convergent-Axis Stereo Synthesis (CASS) Algorithm: Systemic constraints are mapped into unique stereo geometries using an iterative minimum area difference technique.	93
5.6 Convergent Stereo Model: A typical convergent stereo model relies on pure angular resolution to provide accurate depth measurements when the real problem is the integrity of the geometry.	97
6.1 Stereo Two-Planes Calibration: Depicting imaging vectors from each stereo camera pair, and their corresponding intersections with the calibration planes.	100
6.2 Normalized Error Residue vs. Interpolation Order: Three different physical calibration planes were implemented with standoff distances of 124.46, 83.82, and 45.72 cm.	105
6.3 Basic Geometry: Stereo vision system.	114
6.4 Complete Geometrical Description: Solution to point of minimum distance stereo triangulation problem via two planes calibration procedure.	116

FIGURE	PAGE
6.5 Percent Residual Point Error: Error measurements are obtained from the application of calibration points the stereo equations. Measurements are displayed for lower half of the image plane at a standoff of 45.72 cm.	120
6.6 Percent Residual Point Error: Measurements taken from lower half of the image plane at a standoff of 122.5cm.	120
6.7 Percent Residual Point Error: Upper half of imaging plane at 45.72cm.	121
6.8 Percent Residual Point Error: Upper half of imaging plane at 122.5 cm.	121
6.9 Sensitivity of Stereo Equations: Uniform Perturbations in $\vec{P}_{11}$ derived from normal analysis of monocular residual error.	122
6.10 Sensitivity of Stereo Equations: Uniform Perturbations in $\vec{P}_{12}$ derived from normal analysis of monocular residual error.	122
6.11 Sensitivity of Stereo Equations: Uniform Perturbations in $\vec{P}_{21}$ derived from normal analysis of monocular residual error.	123
6.12 Sensitivity of Stereo Equations: Uniform Perturbations in $\vec{P}_{22}$ derived from normal analysis of monocular residual error.	123
7.1 A Perfect Ellipse: Parameterization generates two one dimensional data sets by scanning the x and y coordinates into individual vectors. .	133
7.2 Parameterization of a Perfect Ellipse: Discretization of the input curves causes high frequency noise to be imposed onto the underlying fundamental sinusoids.	134

- 7.3 **Sampling vs Order Grid:** Rows [2 – 6] of the figure matrix have constant order [1 – 5]. Columns of the figure matrix have constant sampling rates [2, 4, ...10, 12]. The first row contains the original image followed by the intersection of it with the ideal PEFR with 500 samples and orders [1 – 5], respectively. 138
- 8.1 **Sequential Block Algorithm:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system. 147
- 8.2 **A “General” Algorithm:** A simplistic model of algorithm design where an algorithm has been tuned for a particular class of input data. If it is used in the classification of a different class of data, performance will suffer. 149
- 8.3 **SBA Network Algorithm:** A parallel model of algorithm design where an individual elements of a process matrix have been tuned for a particular class of input data. 152
- 8.4 **SBA Fully Connected Network:** A more complex model of algorithm design where each data set is input to each process and the results area encoded into a model matrix. 153
- 8.5 **A BAND Network:** A more complex model of algorithm design where each data set is input to each process and the results area encoded into a model matrix which is evaluated and used to adjust the optimal output. 154
- 8.6 **Simple Square Wave:** shown with and without the application of uniformly distributed noise. 161
- 8.7 **Unit Area Square Wave:** contaminated with uniformly distributed noise in the range [-1,1] with BAND and MAP estimates overlayed. . . . 162

FIGURE	PAGE
8.8 Unit Area Square Wave: contaminated with uniformly distributed noise in the range [-1.6, 1.6] with BAND and MAP estimates overlayed.	163
8.9 Unit Area Square Wave: contaminated with uniformly distributed noise in the range [-2.5,2.5] with BAND and MAP estimates overlayed. .	164
9.1 Stereo Platform Design: Stereo vision system designed using the PASS algorithm.	166
9.2 Stereo Platform Design: Stereo system mounted on the Dual Arm Work Module (DAWM) at the Oak Ridge National Laboratory.	167
9.3 Sensitivity Analysis: Estimated range error for object distances of 100 in. with baseline and and focal length perturbations.	169
9.4 Sensitivity Analysis: Estimated range error for object distances of 100 in. with baseline and disparity perturbations.	169
9.5 Sensitivity Analysis: Estimated range error for object distances of 150 in. with baseline and focal length perturbations.	170
9.6 Sensitivity Analysis: Estimated range error for object distances of 150 in. with baseline and disparity perturbations.	170
9.7 Robotic Cutting Task: Stereo vision system designed using the PASS algorithm is providing three-dimensional coordinates that direct the hydraulic cutting tool to its target.	173
10.1 Platform Concept: Front, top, and side views of a conceptual platform with vergence and dynamic baseline capability.	176
10.2 CASS Realtime Computing Engine: A VME-based computing architecture for the implementation of low-level processing on convergent stereo systems.	177

FIGURE	PAGE
10.3 CASS Auto-Focus Engine: Regional Focusing via the Tenengrad function at (5 Hz).	180
10.4 Auto-Focused Image: Toy truck with the region of interest centered on the cross.	181
10.5 CASS Auto-Aperture Concept: Minimizing the difference between the current histogram and a uniform histogram given the input image. .	183
10.6 Two Different Aperture Settings: Illuminance is proportional to the square of the diameter.	185
10.7 Auto-Aperturing Algorithm: The mean of the histogram is used to find a fast, analytic solution to the optimal aperture position.	186
10.8 Corrected Aperture Setting (Too Dark): Aperture position is solved using a weakly calibrated model to give fast performance.	187
10.9 Two Different Aperture Settings (Too Bright): Aperture position is solved using a weakly calibrated model to give fast performance.	188
10.10 Stereo Target Grid: for the convergent stereo vision system.	189
10.11 CASS Platform Design: Design sample without linear actuator to supply dynamic baseline capability.	191
11.1 Block Diagram of Contouring Algorithm:	194
11.2 Volume Segmentation of PET Image Sequence:	196
11.3 Tissue Density Model for the Human Head: The view is transaxial.	197
11.4 Iterative Contour Generation Algorithm:	202
11.5 Volume Synthesis Algorithm:	204
11.6 Attenuation Sinogram Generation:	206

11.7 Uncorrected Image Data: A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.	208
11.8 Synthetic Contour Generation: A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.	209
11.9 Synthetic Tissue Density Map: A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.	210
11.10 Tissue Density Volume View: A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.	211

CHAPTER 1

Introduction

This document outlines some deficiencies in current "state-of-the-art" techniques in passive/active vision. It begins by highlighting fundamental areas in the "black-box", or generic view of vision, that limit the ability to perform robustly across a wide variety of scenes. Specifically, it addresses three fundamental problems in sensing that have plagued machine vision since its inception. Presented generally, they are (1) *Sensor Geometry*, (2) *Calibration*, and (3) *Representation*. Sensor geometry addresses the formulation of optimal sensor configurations given a set of constraints. This will be used in two ways: (a) "off-line" generation of static vision (stereo) systems that will perform optimally in the sense of accuracy under those constraints; and (b) "on-line" configuration of active systems that will dynamically structure the opto-mechanical configuration of the platform. Calibration will focus on the development of an intelligent calibration module that will balance competing precision and efficiency requirements in terms of task-based constraints. A dynamic representational scheme will offer flexible representations that make explicit task-critical data and discard unnecessary or redundant information. Systemic development will concentrate on generating specific applications from user-defined constraints. In particular, a static platform for robotic manipulation tasks and an active platform for object-localization will be developed as "proof-of-concept" using this design methodology. In addition, a theory of system design based on Bayesian networks of competing processes will be used to implement an

algorithm for feature extraction in positron emission tomography of cranial images.

While these problems are common in general vision, they are studied in the context of an application. These applications are specific and closely tied to a set of problems for which they were intended to solve. However, study of these applications is posed in a general theoretical framework that applies to all task-driven problems. Using this road map, outlined in chapter 2, it is possible to use a high-level theory to impel a specific application. Typically there is some separation between, theory and implementation. It is the goal of this work to integrate theory and implementation with a formalism based on task-based knowledge.

1.1 Research Domain

Until recently the theory concerning the acquisition of image data from a scene centered on the production of the "black-box" device. Typically, this device would view a static scene and produce a two-dimensional spatial array (or an image) of the form $f(x, y) = z$, where z represented scene characteristics such as intensity or range. Figure 1.1 depicts the standardized nature of the approach and its direct relationship with the digital computer. The general idea was that anyone could use this device for their application and interact with a familiar representation, the two-dimensional image, to solve a particular problem. Unfortunately the impetus for this idea was not good science; rather, it was motivated by the desire to sell a large number of imaging devices that could work over a broad range of application domains. For a while, the inadequacy of the approach was over-shadowed by the out growth of research these cheaper sensors stimulated in the machine vision community.

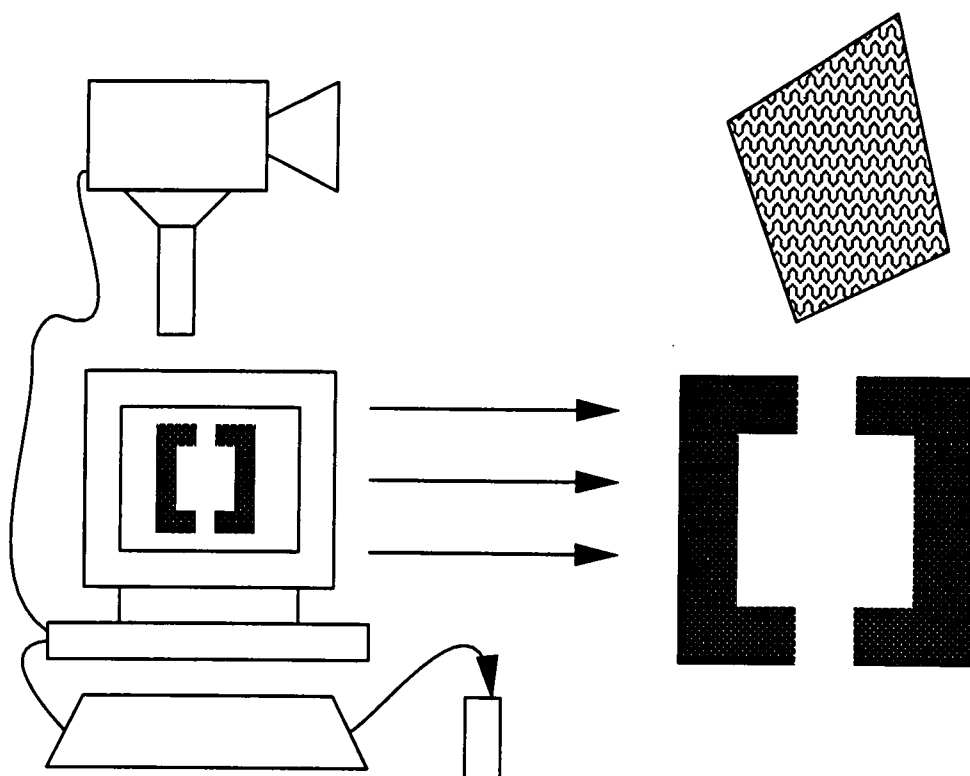


Figure 1.1: **“Black Box” Vision:** A generic model of imaging where an unstructured scene is imaged with a standard ranging system.

Problems with the black-box model began to mount as the research community began to use these sensors outside the range of tasks for which they were suited. Many of these applications had highly specific needs that a generic class of devices, which were designed to exhibit an average set of properties, could not support. These researchers resorted to custom sensors to provide the characteristics that their applications required. In this work, three fundamental characteristics that all imaging devices possess are examined. These are sensor geometry, calibration, and representation. It is the fundamental hypothesis of this work that no single combination of these characteristics or properties can adequately provide useful sensory data for all possible tasks. A corol-

lary to this hypothesis states that given a single task, a particular combination of these parameters can be considered optimal under that task. Figure 1.2 provides a simple illustration of a three dimensional design space that is constrained to operate only in a minute section of the possible design configurations.

The principle problem with this framework arises from the task-domain formulation of the design problem. In other words, given a task T , how does one arrive at a design D ? A solution can be found if the task T is translated into a set of constraints, $C_i \in C$, which have direct correspondence with the mathematical models of the design. If this can be successfully accomplished, the overall design problem can be studied using numerical optimization tools. The structure of this problem can be related mathematically as

$$D = \min f(G, R, M, C). \quad (1.1)$$

where G is a set of variables that represent the geometrical configuration of the system, R is set of properties that describe the desired state of the representation, M is the calibration model, and f is the associated cost function. Locating a stable global minimum in the function f should offer the optimal design, D , for task T . Under this type of systemic analysis, the entire system configuration is derived from a single composite cost function that “best” describes the goals of the current task.

1.2 Problem Domain

The most important question in designing a vision system to paraphrase Jain [Jai94] is “has the problem been defined precisely?” In our minds this definition demands a thorough analysis of task along with integration of this information into machine vision theory. This entails answering questions involving optimal choices of sensors, resolution

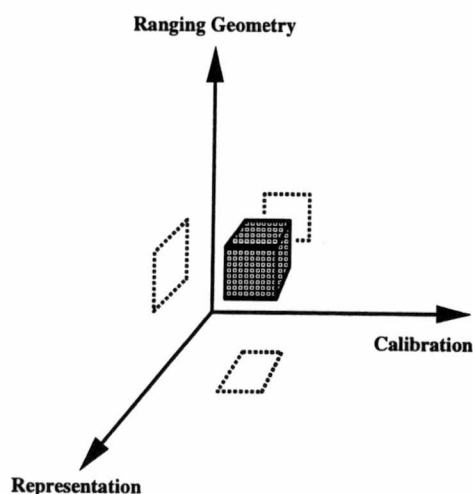


Figure 1.2: **Abstracted Design Space:** A single design cannot possibly represent all possible applications; therefore, design parameters must be tuned for each specific application.

in the data, and computational constraints. Once this type of analysis is complete, then we must ask ourselves in the context of task “What internal representations do we need to adequately perform our task?”. We stress here that this may or may not include recovery as a goal. However, we do agree with Aloimonos [Alo90a] that recovery may be unnecessary for our task, and with Edelman [Ede94] when he says that “reconstruction per se contributes nothing toward interpretation”, nor do we demand the presence of representation. If our task is best performed using the reflexive architecture of Brooks [Bro91] then we will use it. Correctly deciding on representational forms or lack thereof is critical. It must be done in an intelligent, informed, and unbiased manner. As a designer, one should not care if a system is active, passive, or purposive until an objective basis has been formulated for that decision.

In some circumstances, it will be desirable to examine each of the three sets of

design variables in an individual framework. Here the motivation would be to modify the structure of an existing design to perform in a more efficient manner or to find an optimal solution to a sub-problem of range sensor design. In particular three specific class problems are addressed in this work. They are *Stereo Vision Range Sensor Geometry*, *Stereo Calibration*, and *Representation for Identification/Localization*. These highly specific class problems are examined in light of a theory of task domain design outlined in Chapter 2. Using this theory it is possible to design algorithms for design, and to implement dynamical systems using a robust task-dependent engine. Before proceeding to the theory the general issues and structure of the specific problem domains are outlined.

1.2.1 Sensor Geometry

Most range sensors can be classified as *time of flight* sensors or *triangulation* sensors [Jar83]. Time of flight sensors measure distance based on the time difference between the emission and detection of a signal [FGL87]. However, in this work, the ranging device is a stereo vision system using triangulation. Ranging devices based on triangulation use a calibrated geometrical arrangement to solve for three-dimensional locations spatially, i.e. in the case of stereo the intersection of the projection vectors in the camera pair. The numerical precision of the resultant measurements are directly related to the integrity of that geometry. Figure 1.3 depicts the basic geometric structure of the triangulation problem.

Given this model of ranging geometry and assuming that it applies to both *parallel-axis* and *convergent* stereo implementations, the question becomes "How does one decide what geometry is suitable for the given task, **T**." Numerical optimization tools are used

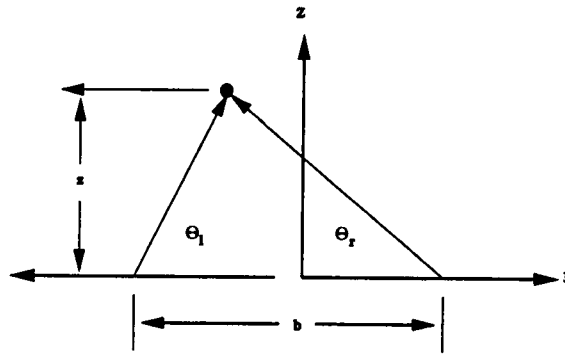


Figure 1.3: **Simple Triangulation:** Three-dimensional depth is solved by the intersection of two measurement vectors. The accuracy of the resultant depth is a function of this geometry.

to provide the solution. In its simplest form, the mathematical structure of the problem is similar to the overall design problem presented in the previous section.

$$D_g = \min g(\theta_l, \theta_r, b; C_g). \quad (1.2)$$

D_g is the geometrical structure of the design; θ_l , θ_r , and b are the represent a base set of design variables; and C_g is a set of constraints derived from basic assumptions about the nature of task **T**. Solution of this problem will result in an optimal geometrical configuration for task **T**.

Applications of this method will facilitate the design of stereo vision systems that require strict adherence to requirements on competing design characteristics such as accuracy, field-of-view, and minimum work space requirements. In particular the utility of this theory will be demonstrated as it is used to develop two fundamentally different types of vision systems.

1.2.2 Calibration

Calibration involves the process of assuming a model for a physical process in a component and using observed data and associated mathematical techniques to find the parameters of the model that give the “best-fit” to the observed data. In stereo vision, the CCD camera is the primary component that requires calibration. And the physical process that needs to be modeled is that of *imaging*. Image and three-dimensional “target” coordinates usually comprise the observed data. A gamut of mathematical techniques exist for fitting models to data which vary to a large extent in accuracy, computational efficiency, and applicability. Also, closely coupled with the concept of calibration is that of measurement uncertainty. This quantity has been viewed as a metric of the effectiveness of a calibration.

If this definition of calibration is accepted, then the classification of camera calibration techniques can be classified in the context of its component stages. Techniques can be sorted in terms of (1) *the structure of assumed models*, (2) *the form, or representation, of the observed data*, (3) *the nature of the mathematical techniques used in the fitting process*, and (4) *the level of uncertainty in the overall calibration*. It must be stressed that quality of calibration can bear no direct relationship with any of these characteristics. Many times, a low level of measurement uncertainty is associated with a “good” calibration; however, many applications of camera calibration may require lower-levels of accuracy. In some cases, the computational burden associated with precise calibration may be catastrophic in the context of a specific application. In any case, it can be assumed that a calibration has a level of measurement uncertainty that limits its applicability in some fashion.

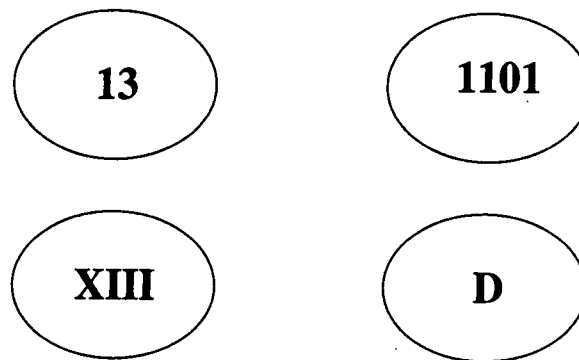


Figure 1.4: **Representations of Thirteen:** Different representations make certain types of information explicit at the expense of other information.

1.2.3 Representation

Marr [Mar82] defines a representation as

“a formal system for making explicit certain entities or types of information together with a specification of how the system does this.”

He goes on to further say that the explicit information is obtained at the expense of other types of information that are obscured and may be unrecoverable. Figure 1.4 shows a simple illustration of this type of phenomena. Thus, in a vision system the choice of representation can have dramatic consequences. Many different views on the role of representation exist in vision [Bro91, MH80, Alo88]. Perhaps the most hotly debated representation scheme is *recovery*. This paradigm tries to reconstruct as much of the physical detail from the scene as possible. The advantages and disadvantages of this scheme are clear. In defense of recovery, it has the desirable characteristic of making no information explicit at the expense of other information. It, however, does not aid the inference process and it may not even be necessary or efficient in the production of robotic vision systems. Some researchers such as Rodney Brooks [Bro91] claim that

representations are not required in any fashion. Intelligence, in his mind, can emerge from the collaborative effort of reactive systems. Aloimonos [Alo93a], on the other hand, claims that representations should be minimal and a function of task. Of course, reconstructionist such as Tarr and Black [TB94a, TB94b] demand strict adherence to the recovery paradigm.

Jarvis [Jar83] assumes "the type of range finding required is that which results in a rangepic, an array of distance estimates from a known point or plane with adjacency constraints corresponding to those of two dimensional imagery". Besl [Bes88] states that a range image "is a large collection of distance measurements from a known reference coordinate system to surface points on object(s) in a scene". The range image is the standard representation that most stereo ranging systems build for interpretation by computer vision systems. It has been used due to its versatility and direct spatial relationship with intensity imagery. New types of representations that have direct relationships with the application task are desired and may prove justifiable both computationally and qualitatively.

1.3 Overview of Research Contributions

As outlined, the contributions of this work are focused into three related areas: (1) optimal geometric configuration of active/passive stereo ranging platforms, (2) development of calibration models for stereo ranging, and (3) intelligent representations for task domain stereo ranging. In particular, some unique qualities of this work have already been recognized by the machine vision community. They are:

- An algorithm for analysis and synthesis of calibration polynomials for stereo camera calibration [TT95b].
- An algorithm for the design of parallel-axis stereo systems [TT95a].
- An algorithm for the design/configuration of convergent-axis stereo system [TT95a].
- A robotic vision system for small object manipulation [Tha93, TT95a]

1.4 Organization

This text is separated into distinct parts with a common goal. Chapter 2 provides an outline of the step-by-step guidelines that were used to implement each of the three application specific problems. Using these guidelines it is possible to formulate optimal designs to task-driven problems. Chapter 3 constitutes a brief review of the concepts and traditional platforms of active photometric stereo. The purpose is twofold. First we would like to review one aspect of the history of stereo vision to provide perspective. Also, we would like to evaluate active stereo platforms in the context of kinematic and optical capability. Chapter 4 develops the fundamental mathematical tools for the construction of static-baseline binocular stereo platforms. Chapter 5 illustrates the principles of designing convergent stereo systems with dynamic baseline configurations. Chapter 6 provides a system oriented view of the calibration of a pair of stereo cameras. Its advantages and disadvantages are outlined and discussed in terms of computational efficiency and accuracy. Chapter 7 looks at the basic issues in the development of representations for task domain ranging. These representations should

maintain a direct relationship with the desired perceptual tasks and will facilitate interface with perception, learning, and behavioral mechanisms. In Chapter 8, a high-level design methodology for purposive vision systems in a modified Bayesian framework is presented. Using this theory, it becomes possible to produce robust and accurate vision algorithms together with a reduction of complexity in the components of that system.

In addition to the basic theory as presented in the main body of this document, a series of applications is provided that document the implementation of the basic theory into three distinct computer vision applications. Chapter 9 discusses the implementation of a passive stereo vision system for small object manipulation. This system has strict requirements on the accuracy of range estimates. Chapter 10 relates the implementation details of a system for small object localization. This system was designed to be highly dynamic and loosely structured for object identification. Lastly, Chapter 11 used the Bayesian network to implement a highly task specific application, the extraction of brain contours from positron emission tomography. While it is a departure from the basic theme of stereo vision, it provides perspective on the applicability of this fundamental theory to other modalities of vision processing. Finally, the conclusions and overall results are discussed in Chapter 12.

CHAPTER 2

A Task-Oriented Design Methodology

The impetus for the design of any vision system is to do something. However, in many philosophies of general vision [Mar82] this important piece of information is completely ignored; implementation is left as a detail. While this type of approach has led to many important advances in machine vision theory, it has somewhat hampered the production of robust, stable, useful vision whose primary mission is the completion a specific task.

Many researchers have examined the concept of integrating task information into various components of a visual system [Alo88, Bro91]. This important work has reset the stage for a holistic design view of vision, where task-level information is considered mandatory in algorithms, as well as implementation details. However, the primary criticism of this school of researchers is they have seemingly rendered machine vision to a mosaic of “nuts and bolts”, bottom-up approaches that don’t seem to have the strong theoretical or analytic background that has served the development of Marr’s vision so well.

It is the focus of this work to provide a holistic framework for task-level design of visual systems. It is our intention to isolate a particular task and its constraints, e.g. accuracy, work-volume, execution-speed, etc., and integrate these constraints into each component of system design. Thus, one does not look at task-level vision as a methodology for the down-stream processing of raw data; rather, it is process of *optimizing*

at each processing stage the data-stream to simplify and stabilize the task set. In some cases, task-domain vision has been classified as an algorithmic approach [TB94a]. It is our hypothesis that these statements are totally false. The central thesis of this work is: **Task-domain vision is the isolation of sensory components, calibration of these components, representations of the sensor data, algorithms that operate on combinations of these, and a set of constraints that governs the interactions of these components with the task and the environment to produce vision systems that perform a specific task in an optimal fashion.**

2.1 Pruning the Possibility Tree

One problem with the undertaking of attempting to attach a high-level perspective to the problem of "bottom-up" design is how to avoid becoming unavoidably attached to the task. In other words, how does one separate the principal from the overtly practical. This work attempts to frame the concept of task-specific design into a type of constrained optimization problem. Using this paradigm, it is possible to use a general framework for constructing visual systems, while at the same time isolating task critical knowledge into convenient representations that seamlessly integrate into our design structure. In this way, a system design begins with general theoretical analysis; proceeds to the derivation of critical task components; and finally it culminates in the use of optimization theory to produce optimal task design.

While this type of theory is ideally suited for global system design, it also has merits in the design of sub-optimal systems, where a single component of a more general system is designed with purposive flavor and the rest of the system formulated from a more

traditional viewpoint. Indeed, this type of theory has found many applications in the mathematical theory of optimization and the formulation of sub-optimal solutions to problems [Str86, Str80, RRR83]. Thus the procedure as hinted to in the introduction follows:

- MODELing the system: (MODEL)
- Task Requirements And Integration of kNowledge: (TRAIN)
- FOrmulation of Cost functions Using Specification : (FOCUS)
- Formal IteRative Evaluation: (FIRE)

2.2 MODEL

Modeling is a common process that involves examining the dynamics of system design and encoding these dynamics into a mathematical structure where state variables represent certain configurations of the system. In a sense this process is a mathematical description of the design space where the equations relate how state variables interact to form a complex system. Modeling is more of an art form than a science and requires excellent judgment on the part of the implementor. The user must model enough of the system behavior to cover the necessary regions of the design space as specified in the system requirements; however, the user must not try to implement overly complex or unstable models which do not lend themselves well to optimization.

As in the initial discussion from Chapter 1, a part of the design equation 1.1 was the sensor model, G , and calibration model, M , and a representational model, R . These items can be thought of as component modeling, while integration of these mathematical

models can be a global system model. Either *component modeling* or *global modeling* is acceptable to the TODEM methodology. Isolation of either the global model or the component model implements an infinite design space as discussed in following chapters. For instance, a ten variable global state model requires choosing a design from a ten dimensional infinite design space! It may be nearly impossible to pick an optimal design using intuition or trial and error methods. It is the purpose of the next three stages of the TODEM framework to outline a procedure for iterative recovery of the optimal design for a task, **T**, from such oppressive design spaces.

2.3 TRAIN

A prudent step in the pursuit of any design effort is the specification of formal requirements for a system. In vision, the requirements are usually in the form of some sort of accuracy goals together with some physical limits imposed on the design through the use of "off-the-shelf" components or the working environment. Usually this type of specification is rather cursory and ad hoc in nature. In the TODEM system, it is rigidly enforced.

In formal design, a list of user and developer requirements is commonplace. Therefore, it is a worthwhile endeavor to incorporate this sort of familiar worksheet into vision system design. In chapter one, it was stated in Equation 1.1. that a design **D** consists of a set of variables, **G**, that represent the sensor configuration; a model of the calibration, **M**; a mathematical description of the representation, **R**; and a set of constraints, **C**, that encode rules of interaction of each of these components. It is the purpose of the **TRAIN** step to isolate task requirements and integrate these requirements into a

system of constraints that bound either the global and component design spaces.

The constraint superset can be viewed by the composite,

$$C = \{C_e, C_a, C_g, C_m, C_r, C_t\}, \quad (2.1)$$

where C_e is a set of constraints derived from the environment, C_a is a set of artificial constraints such as regularization parameters, C_g describes sensory geometry boundaries and constraints, C_m is the calibration model limits, C_r is a set of representational constraints, and C_t is the list of constraints that are particular to specific task level activities.

Environmental constraints are physical constraints that are derived from important characteristics of the local workspace. Examples of these types of constraints are space and volume limits, radiation exposure levels, lighting conditions, local surface composition, etc. Identification of these characteristics is extremely important in the initial stages of design. Sensory geometry constraints represent boundaries that are derived from the platform or housing of the sensor, e.g. local space and volume limits, power consumption, proximity limits, and other sensory design assumptions. These characteristics represent the interaction of the sensor with its local environment such as a mobile robot or manipulator arm. Calibration model constraints identify unique features of the calibration such as accuracy and precision requirements, computational loads, mathematical stability, etc. These parameters are important in the derivation of information from the scene that is tuned to perform a specific task. Representational constraints are similar to calibration constraints in that they impose restrictions on low-level processing to produce timely and efficient data with respect to task requirements of the algorithmic components of a device. Artificial constraints can represent filters or other regularization steps that help the interactive design equations converge to a stable place

in the overall design space. Finally, task level constraints can be used to represent constraints to a system that is responsible for multi-task behavior. Thus a hierarchy of task behaviors can be stored to facilitate operation in dynamic environments.

2.4 FOCUS

Alone a set of constraints and a process model do not solve anything. A set of goals together with these constrained model can be mapped into a mathematical formalism that if properly conditioned will produce an optimal design. In optimization theory, the form of these goals is usually a cost function. Thus, using the system goals and measurable characteristics of the design, an optimal design is achievable.

The cost functions can take two forms: (1) a global measure that takes accounts for "cross-talk" between system components; and (2) a separable, or component-based, model that assumes each system module is independent and can be studied individually. Global modeling is much more demanding and can often lead to complex cost functions that do not lend themselves to iterative optimization. However, it has the advantages of expressing all the system goals into a single mathematical framework that will deliver a more optimal solution. Component modeling is simple, and will tend to converge to a set of state variables more rapidly, also in component-based models analytic solutions are more likely due to reduced complexity of the mathematical models. Once the FOCUS step has produced a satisfactory cost function, the state model together with the integrated constraints is delivered to the iterative optimization procedure to produce an optimal configuration of the model state.

2.5 FIRE

Most complex designs do not have analytic solutions that are readily obtained. Iterative optimization is a simple mathematical tool to perform complex and multi-dimensional searches. It should also be stressed that although the iterative approach is the most general, it is by no means required for the TODEM step. If the problem is formulated in such a way that the constraints reduce the design space to a single point, an analytic solution is possible. Then, the final step is simply solving for the unique configuration of state variables that satisfy the design equations.

2.6 Conclusions

A simple procedure for organizing and implementing the design of task-domain systems, TODEM, has been outlined in four steps. The first step is a simple state modeling of the possible design space. Each of the design parameters is encoded into a variable that serves together with the constraint equations derived in the TRAIN step to define a bounded design space. This bounded design space serves as a domain of support for the cost function derived in the FOCUS step. Here, design goals are represented by cost functions which are a convenient forms for encoding conflicting system goals and uncertainty relationships existing in any practical design. In the final, FIRE, stage of the process, a constrained iterative optimization is performed over the design space to produce a set of state variables that offer optimal performance for the task goals. The procedure can also be used to study dynamical problems where certain state variables from the system model need to be reconfigured when relevant task information changes due to changing environment conditions or new information is presented to the sys-

tem. Thus, dynamic scene conditions are simply handled by repetition of the FIRE stages where state variables are directly linked to things such as motor speeds, actuator distances, and processing rates.

The four steps of task-based design are used to guide the designs of three systems that are from an applications domain. In the first case, a global model is used to design and calibrate a passive stereo vision platform for small pipe localization. The next system, uses a global model to design and calibrate an active stereo vision system for rapid object identification; then, a component model is used to offer new state configurations based on relevant task information such as changing lighting conditions and dynamic object distances. Finally, a system of object localization and identification is presented. This system used a component-based model to design a purposive algorithm for the extraction of scalp and brain contours from computed tomography images. It is shown in detail how each of these systems benefited from the TODEM design methodology to produce optimal system designs with respect to their individual assumptions.

CHAPTER 3

Literature Review

3.1 Introduction

Vision, in general, provides a mechanism for obtaining information on the current state of the environment. More specifically, general vision occurs in two steps: (1) *Sensation* and (2) *Perception*. Sensation is the physical process of registering forms of energy from the environment and transforming them into an internal representation of that energy. In biological vision, sensation happens as visible radiation is converted into electrical signals. In computer vision, sensation typically results in an image grid quantized both spatially and in amplitude. After sensation occurs, the perceptual process involves operating on that representation and producing some numeric or symbolic constructs that are correlated with the status of environment as it was recorded in the current sensation frame. Figure 3.1 depicts the basic sensation-perception cycle. This has been the traditional conceptualization in the machine vision community for some time. Abstracting vision in this nature evolved quite naturally from the early applications of machine vision [Fre74, TB76]. Scenes were typically composed of static objects imaged under controlled lighting. Under these circumstances, most problems could be attacked with this idealistic methodology. The sensation parameters were controlled and fixed previous to the perceptive process.

Recently, a number of researchers have become dissatisfied with the classical view of machine vision. They site the inability of classical applications to perform robustly

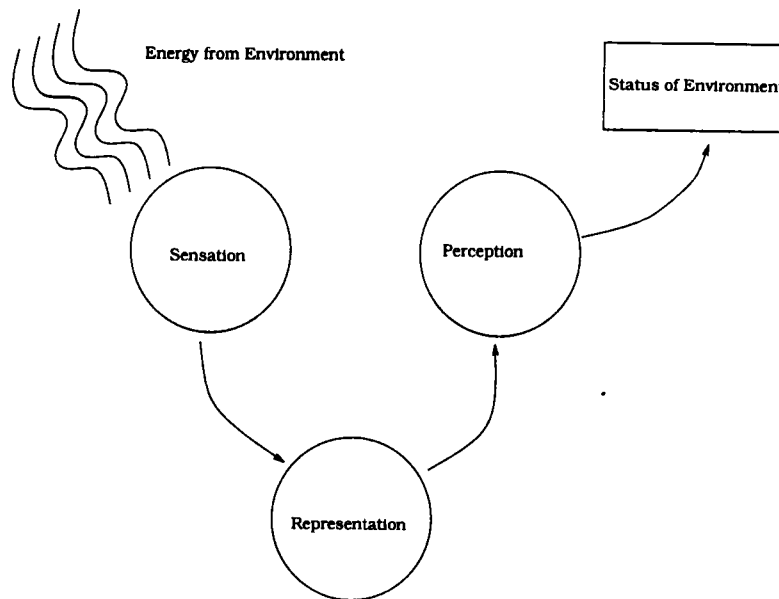


Figure 3.1: **Passive Sensory/Perception Cycle:** The cycle is strictly open-loop. And, the sensation and perception modules have been abstracted into two separable components.

when faced with dynamic changes in scene conditions and the intractability of many classical vision problems [Alo93b]. Arguing from a biological perspective, they claim that perception is inextricably linked with sensation. In fact, the sensation process must be intelligently controlled by feed-back from perceptual information. Figure 3.2 represents the strong linkage between sensation and perception. Although, the addition of a link between sensation and perception is conceptually a simple one. It has been shown to have dramatic effects in the composition of biological vision. In humans, for instance, the sensory apparatus has several degrees-of-freedom (DOF) (optical and mechanical). If ability of the subject to effectively exploit the full range of this freedom is hampered, then the quality of perceptual data can be severely degraded [Gib89]

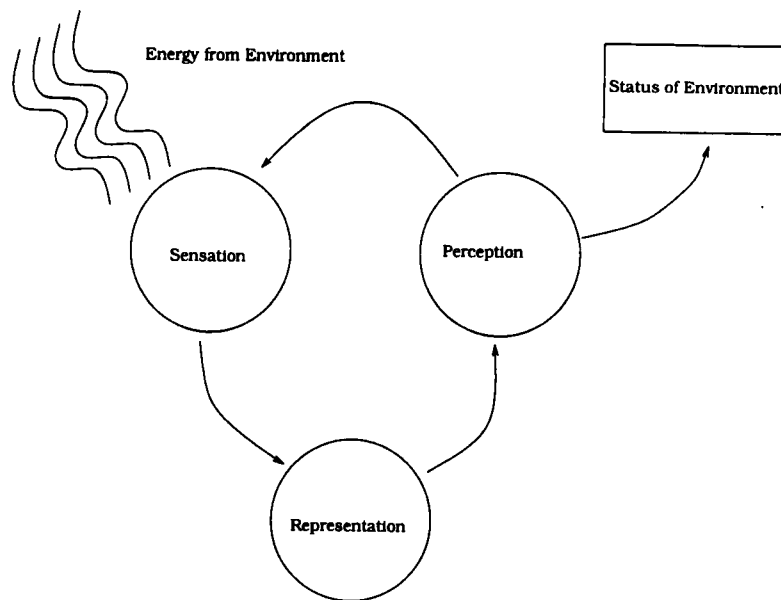


Figure 3.2: **Active Sensory/Perception Cycle:** The cycle is strictly closed-loop. And, control of sensation is executed by the perception module to provide intelligence to the sensation process.

3.2 The Active Paradigm

The philosophical and scientific debate concerning the phenomena of perception as an active or passive process, as previously outlined, can be traced in the philosophical context to Goethe [Goe67] and in the scientific to Gibson [Gib89]. Both men argued that perception was trivially or axiomatically active, i. e., the action of the machinery of perception could not be separated from the process. Everyone who has worked with real imagery implicitly understands this principle as much time is spent tuning hardware parameters to give optimal image data points. Ikeuchi [IH90] concurs saying, “The choice of vision components (representations, features, sensors, etc) governs the quality of the resulting vision programs”. However, this was a clear contradiction of Marr’s [MH80]

modular hierarchy of vision where implementation was considered a low-level detail that could be excluded from the perceptive theory. In the early nineteen-eighties, the inability of the applications of this theory to produce even moderately robust machine vision systems was the source of rising discontent among many researchers. The source of this discontent was the alleged inadequacy of his theory in providing a basis for observer interaction. In fact many vision researchers, in particular Aloimonos [Alo93b] claim that Marr's concept of the existence of a general theory of vision where no assumptions need be made and recovery of information can be total is an impossibility.

Development of active, or animate [Bal91, BB93], vision as a viable paradigm in computer vision has been most associated with the work of Bajcsy and Krotkov [Baj88b, KFS88, Kro89] at the University of Pennsylvania. Under this renewed methodology, a theory of vision must be integrated with the machinery of vision. Bajcsy in [Baj88a] says that "perceptual activity is exploratory, probing, searching; precepts do not simply fall onto sensors as rain falls onto ground". According to Pahlavan [PUE93], "the observer selectively acquires visual data in space and time". Sandini [SGGT93] relates, "that simultaneous and coordinated operation of vision and action can be used not only to simplify some traditional visual tasks but also to extend the overall scope of vision to important new areas". And Aloimonos [Alo90b] reveals that, "We cannot study the general problem of vision, not tied to any applications". These statements from respected researchers in active vision outline some of the new concepts in vision that are very exciting.

3.2.1 Concept I: Selective Data Processing

Swain and Stricker [SS93] have characterized the typical digital image as follows:

- “It contains far more data that can be analyzed by a practical vision system in real-time.”
- “It is highly redundant”
- “Information of importance to a given visual task is episodic within the source ...”

If one accepts these generalizations, then the need for discriminate processing within the image space becomes evident. Active vision address this problem in two ways. First, using a model derived from biological vision systems that use a single high resolution region of the image in conjunction with a “stack”, or pyramid, of lower resolution regions that compose the remainder of the image space [Bol89]. Using *foveal* techniques, DePiero [DNG92] reports bandwidth reductions in a video transmission system for remote driving of 1000 : 1. While Swain [SS93] that foveation on the order of the human visual system could result in a bandwidth reduction of 10000 : 1. The second and final stage of discriminate processing is motivated partially as a consequence of the first. The reduction in bandwidth caused by foveation together with the need for sophisticated systems that mimic biological eye movements such as tracking require the existence of intelligent *gaze* mechanisms. Gaze control consists a group of behaviors at the task-level that direct the high-resolution foveal as well as any mechanical degrees-of-freedom to a point interest in the scene. Figure 3.3 shows a structured-lighting application where 99 % of the image contains irrelevant information. Unless intelligent processing techniques are employed, unnecessary computational costs will be incurred.



Figure 3.3: **Selective Data Processing:** Only a portion of this image contains information of interest to the user. (In this case a structured lighting system). Most of the image data can be disregarded. (Courtesy: F. W. Depiero)

3.2.2 Concept II: Exploratory Vision

Ballard [Bal91] predicts that the “central asset of active vision is gaze control.” The mainstay of his argument is built around the assumptions of the salient nature of image data outlined by Swain and Stricker [SS93] and presented in the previous section. In this framework, a vision system with foveation must possess the capability to adequately direct the foveal area. The machinery of this ability is *gaze control*. It can be defined as the use of imaging degrees-of-freedom in the execution of visual tasks. It must be stressed that this use is not liberal or haphazard in application and must be coupled with an important *behavior* present in all biological vision systems - *Attention*. Figure 3.4 represents an information-dense scene. Vision systems must have the ability to explore such scenery in an intelligent and timely manner.

Attention is a process or behavior that directs the hardware of vision to a region/object of interest in the visual space. Aloimonos [HA92] describes this as the exploration of the visual space for activities and finding the one that most facilitates the goal. Brown cautions against the characterization of these of abilities as behaviors implicit in vision [Bro94]. Nevertheless, a mechanism must exist for directing a sensory-motor system to an interesting region in the visual space. Whether attention is behavioral in nature or not is a secondary issue. Attention must overtly present optimal data from an expansive set to the system in a timely and efficient manner. Exploration also demands mechanisms for *Change-of-Attention*. Abilities of this nature can be easily justified in an evolutionary context; however, it is sufficient to say that if a vision system has engaged its attention to region A and after some time information in region B has become relevant to the system at a critical level it would be foolish to complete processing in A. Change-of-attention mechanisms provide this capability.



Figure 3.4: **Exploratory Vision:** An information-dense image. Intelligent systems must have the ability explore these types of scenes in an efficient manner.(S. Dali: April in Paris)

3.2.3 Concept III: Regularization of Vision Problems

Many problems in vision are ill-posed when formulated in the context of Marr's theory. For instance, the correspondence problem in binocular stereo does not have a closed-form solution based only in the confines of computational theory. The reasons for this as illucidated by Aloimonos [Alo88] are

Usually the constraints from the geometry and the physics of the problem
are not enough to guarantee uniqueness of the computed parameters ...

He goes on to state that additional constraints must be imposed to regularize typical problems. The addition of strong constraints which are not represented in a wide array of scenes may lead to instability or failure when applied to real scenes. Machine vision systems constructed under the regularization paradigm are usefully only in tightly controlled industrial environments where the lighting and other scene variables are well-controlled. It is the contention of active vision researchers that their theory can remove the need for regularization by using the knowledge of observer action to make vision problems well-posed. Thus, **Uniqueness** and **Robustness** can be obtained without the mitigating factor of limiting scene dynamics.

3.2.4 Concept IV: Task-Driven Theory

Along with the concepts of blending of theory with implementation, the paradigm of active vision has revitalized an important concept from the early days of vision - *Task-Domain Knowledge*. Purposive Vision [Alo93b] is active vision integrated with task-level knowledge. It is not the overtly specific knowledge criticized by Marr; rather, it is using a priori knowledge to reduce the computational complexity of design. This new ideology

considers vision as a group of processes that solve a visual task. There is no separation of theory and task. Systems are built through a dual analysis of theory and task. In this way, groups of tasks-specific visual systems interact under the control of "intelligent" behaviors resulting in the synergy of complex or sophisticated machine vision.

3.3 Active System Design

In the previous section four major principles that drive the construction of active vision systems were presented. They are:

- Selective Data Processing,
- Exploratory Vision,
- Regularization of Vision Problems, and
- Task-Driven Theory.

While these principles must certainly govern the interplay of vision hardware and vision algorithm in a direct and efficient manner, they also must be applied in the design of active hardware. In doing this, the concept of a "purposive" system is satisfied in its purest form. In this section, we hope to outline the implications each one of these principles has on the design of integrated sensory-motor systems and form a basis for comparison of some of the important work in the development of active vision systems that is present in the vision community.

3.3.1 Characteristic: Foveation

Foveation is a characteristic of active vision that needs to be considered in the hardware design. As mentioned, there are two popular ways of implementing it. The first consists of a software approach as implemented by [Bur88, DNG92, Bol89]. It is primarily a computational technique that consists of partitioning an image into several regions of varying resolution. The highest resolution region thus becomes the "fovea". Another approach that employs a hardware implementation, uses zoom-lens systems to simulate foveation [LL95]. By varying the focal-lengths of the lens system, one aspect of foveation is simulated. In this method, however, data density with respect to the image is not reduced as it would be in a full foveal implementation. Nevertheless, most designers prefer to avoid the computational burden associated with software foveation and employ a zoom-lens system. As a result, provision of foveal capabilities is necessary in the first stages of design.

3.3.2 Characteristic: Gaze Control and Attention

In a foveal system, the ability to direct the visual hardware to a point of attention is critical. This ability is called *Gaze Control*. It can be defined as the act of directing the vision hardware in such a manner as to keep a point/region of interest within the viewing region of the sensory device [Bal91]. In highly dynamic and rugged environments the action of gaze control might also include some type of scene stabilization. Scene stabilization is required by automated vision processes that operate under such conditions to insure that degradation of the results is limited [BA94a, BA94b]. With respect to active platforms, the first aspect of gaze control is the most important. Thus, the potential for gaze can be directly related to kinematic configuration of the robot. Robots

with restricted DOF may be limited in their ability to perform efficient gaze control actions. Robot head design must begin with an understanding of what gaze motions need to be performed and only then can the kinematics of the head can be properly derived.

The most desirable feature of an active robot head is its ability to control its own sensory hardware. Execution of these abilities should occur so as to make perception of the scenery easier [HA92]. Attention has been defined as the preferential allocation of visual processing within a scene [RB90] [WF92]; therefore, it is the process that directs the gaze control to a new fixation point. Although it is also related to the kinematic configuration of the head, it is primarily a software process and will not be discussed in the context of head design.

3.3.3 Characteristic: Integration of Task-Domain Constraints

Most interesting computer vision problems are ill-posed. The standard technique for the solution of ill-posed problems is through the introduction of *a priori* knowledge in the form of constraints. Constraints can be introduced in two ways: (1) Constraints based on the physical nature of the problem (2) Assumptions introduced due to specific goals that the formulation may require. Top-down theories of computer vision [MH80] dictate that the first set of assumptions provide an adequate basis for the construction of visual systems. However, many implementations of top-down theories employ a restrictive set of constraints that result in *ill-conditioned* systems that will fail if applied to a large variety of scenes. The relaxation of physical constraints can be affected by the introduction of task-domain constraints that can be used to configure a design along with associated physical assumptions into a stable state. This correlation of task with

design will result in *well-posed* systems that will be applicable to a wider variety of scene conditions within the limits of the task-domain constraints. Thus, this principle is the single most relevant idea that can be applied to active system design. A design must be configured for a task. If this is done properly, the design equations will occupy a more stable state.

3.4 Active Platforms

Several machine vision laboratories have relocated their activities in passive machine vision to a more active approach. These active systems are characterized by an integrated mechano-optical design philosophy. Construction of vision systems based on this idea begins with the design of robot heads loosely-based on a model of the capabilities of the human head given in Table 3.1. A robot head could be defined as a camera system (usually binocular) together with some form of mechanical positioning device that is capable of directing the focus-of-attention of the camera system. Active systems is algorithmic concept built upon the vision platform- usually a robot head. Pahlavan and Eklund [PE93] state that active vision is active integration, implying that there is more to the design than simply an mechano-optical device feeding a computer and carrying out commands- the degree of integration is paramount. Thus, the design of active systems is not just the design of dynamic camera systems, it is the design of vision algorithms that intelligently manipulate a mechanical apparatus to perform a task. The most common application is the use of motion parameters to simplify traditionally ill-posed machine vision problems.

The study of active systems can be broken into two related areas: First, the con-

Table 3.1: Primary Mechanical Capabilities of the Human Head.

Axis	Range	Resolution	Speed
Pan	± 90.0	NA	NA
Tilt	± 45.0	2.0	800.0
Torsion	± 20.0	NA	NA
Vergence	± 45.0	2.0	2.0
Version	± 45.0	2.0	2.0

struction of robot heads for active vision research. Second, the design of active visual systems that will reside on these heads. It is acknowledged that both of these activities are highly related and one should not be undertaken without considering the other. This section will explore the development of robotic heads that were constructed in several laboratories.

3.4.1 Early Work

Some substantial work was done in the arena of integrated sensory-motor systems before the benefits of active vision were completely understood. These pioneering efforts marked the first uses of computer controlled imaging hardware for use in time-varying scenery or viewing conditions. While many of these systems are significantly different from the current concept of an active robot head, they do embody many of the principles of active system design and they serve as the foundations for the current generations of robot heads that will be discussed in detail.

One of the earliest documented sensory-motor platforms was built at Stanford [Ten71] in 1971. The system consisted of a pan and tilt unit, variable focal-length

lens turret, and wheel mounted color filters. Used in conjunction with Moravec's [Mor81] slider system, it could provide stereo data to the perceptual engine. Waseda University constructed a true stereo system [OKK73] in 1973. The WABOT was capable of pan rotation, convergent rotations, programmable focal length, and scan line selection. In 1983, Carnegie Mellon University built the POPEYE vision system [BSS93]. The POPEYE system was kinematically similar to the Stanford arrangement; however, it offered a tremendous improvement in processing power due to advances in computer and frame grabbing technology. Its main disadvantage was the lack of translational movements necessary for generating stereo data. In 1985, Kuno [KNI⁺85] developed a stereo camera head capable of baseline translation, yaw rotation, and tilt rotation.

Each of these early platforms employs some sort of positioning device in a relatively straight forward manner to offer flexibility to an imaging system. They differ significantly from the current class of heads in that their level of sensory-motor integration is small and their designs were not based on any *anthropomorphic* similarity with the human head. The current generation of robotic heads are distinguished from simple motorized sensors in two specific ways: (1) most are loosely based on the a kinematic model of the human head (2) the perception/sensation loop is intricately coupled.

3.4.2 The Harvard Head

Ferrier and Clark [FC93] describe a system that has the ability to control both the position of the camera (through pan and tilt motions) as well as optical parameters such as focus and aperture. They claim a design based on models of the mammalian oculomotor control systems. Their main interest is centered on the development of oculo-motor control modules for *gaze* and *fixation*. They partition the pan, tilt, and

vergence axes of the head mechanism into three descriptive regimes: Saccades, Pursuit, and Vergence.

Saccades is defined as a rapid motion of the pan, tilt, and optical axes which results in a change in the gaze point of the system. Vergence consists of forcing the optical axes of each camera (for a stereo system) to intersect the gaze point. Finally, pursuit is an action following a saccade-vergence sequence designed to track an object located at the gaze point. While the definitions of these actions has remained fairly constant, implementation details have been contested. In particular, the pan, tilt, and vergence mechanisms of the Harvard Head are based on Robinson's model of the human oculomotor system [Rob91]. This model has two interesting features, it is shown in Figure 3.5.

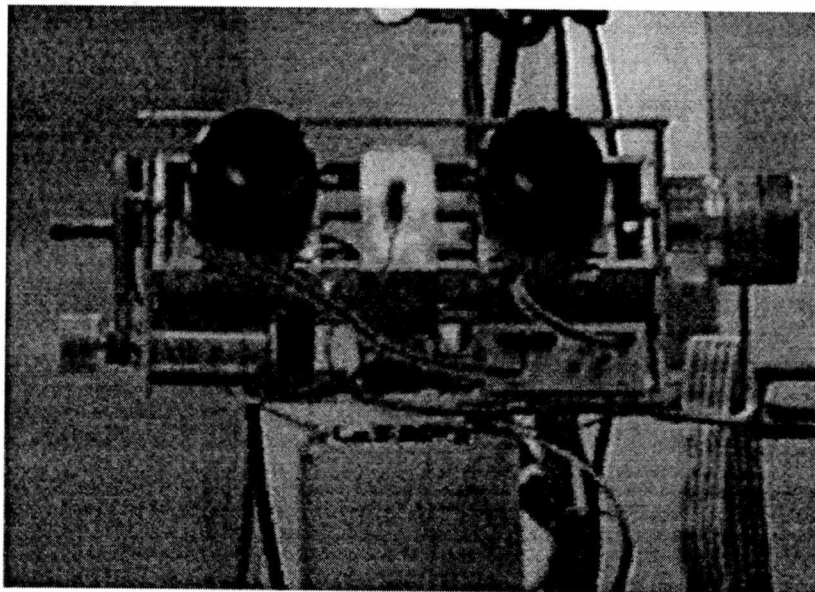


Figure 3.5: **The Harvard Robotic Head:** A low cost design for gaze and fixation experimentation.

- It assumes a sampled data model for the saccadic system.

- It requires positive feedback in the control loop to prevent oscillations caused by time delays in the negative feedback loop.

Young and Stark [YS63] treat the pursuit system as a sampled data system. Ferrier and Clark [FC93] admit that this latter turned out to be untrue - or at least the sampling interval of the pursuit system was much higher than that of the saccadic. Information on the mechanical capabilities of their system is summarized in Table 3.2.

Experimentation

Experiments on the Harvard head were performed in three areas: (1) Tracking, (2) Attentive Servoing, and (3) Depth from Controlled Head Motion. Tracking software was written that followed "blobs" of high intensity on a Datacube system. Both pursuit and saccade movements were implemented and verified. Attentive servoing was a higher level application that tried to locate interesting features in the environment and then fixate on them. The software extracts primitives (line ends, color, orientation, and motion) from the visual field into a set of feature maps. A weighted sum of these feature maps form a saliency map for selection by the system. The most relevant feature to the task is then selected and fixation occurs. Finally, small eye movements were used by the system to compute optical flow. When this information was combined with knowledge of the motion, a depth map was computed.

Discussion

The authors claim that a custom built robot head was designed and tested for less than ten thousand dollars. They demonstrated several basic algorithms on an active system. No comparison was made between the active approach or the common passive

Table 3.2: Primary Mechanical Capabilities of the Harvard Head.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	±130.0	0.036	0.036
Tilt	±35.0	0.036	0.036
Vergence	±15.0	0.036	0.036

techniques. They have suggested several improvements as follows.

- The control system should be more **anthropomorphic**, i. e., it should exhibit more self initiated action.
- The lack of torsional control prevents simplification of the stereo correspondence problem.

3.4.3 The KTH Head

Pahlavan and Eklund [PE93] have designed a 13 Degree-of-Freedom (DOF) head that in principle imitates mammalian heads, shown in Figure 3.6. However, their idea of system design has evolved beyond that of Ferrier and Clark. In particular, they espouse two main principles.

- Active Vision is Active Integration.
- Active Vision is Purposive Vision.

The first principle states that the separation between the visual and mechanical systems must be minimized. This means that algorithms should be designed with the intent of exploiting the mechanical drives in an intelligent way. From principle two, the idea

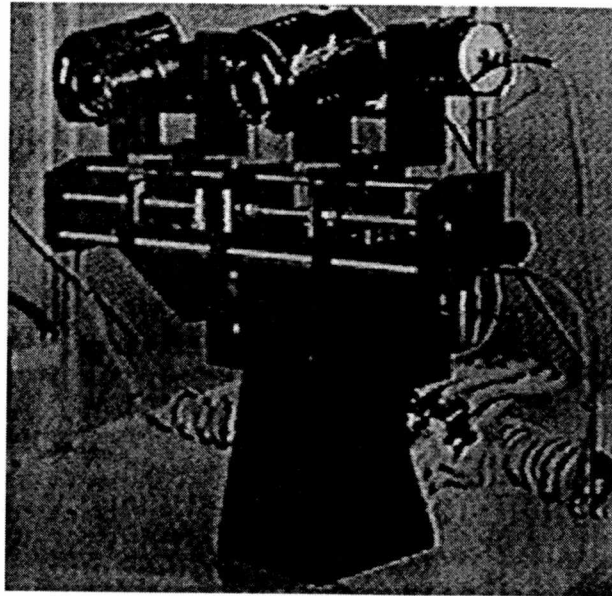


Figure 3.6: **The KTH Robotic Head:** A anthropomorphic design to imitate mammalian head movement capabilities.

that an active vision system can be used for general purpose vision is dispelled. Active systems must include a behavioral engine that directs the head in a purposive manner, i.e., in the context of some task. From here, their ideas transform into some technical issues.

System Design

The authors state that in most head-eye systems there is no distinct separation between eye and head movements. They claim that eye movements do not affect the visual cues, i. e., no geometrical distortion is generated by movements of the eye. They proceed to claim the same is not true for neck movements since they generate a change in the perspective of the images. They conclude with the proposal that previous designs have only studied the geometrical inter-ocular relations rather than imitate mammalian

vision systems. Biological solutions, they claim, serve a fundamental purpose in their overall construction but they do not comment further.

Experimentation

They have constructed a control architecture around a network of transputers which allows them to change all of the mechanical and optical parameters in parallel. At the systems level they have implemented primary ocular reflexes consisting of motion detection, auto-focusing, vergence-pursuit with stabilization, and saccades. These processes are typical of active systems with the exception of stabilization which they define as “nothing but a trigger function of the gaze-holding and pursuit process”. Experimental verification was accomplished using the familiar bright light tracker from which they produce plots correlating motor response to active visual moments. Some of the general capabilities of the head are outline below:

Discussion

A robotic head based on the mammalian visual system with thirteen degrees of freedom was constructed. System design began with the concept of equipping the system with a behavioral engine before it could be a true active system. The authors point out that more inspiration should be gained from biological solutions to the problems. They have suggested the following improvements:

- Implementation of a set of primary reactive behaviors for continuous interaction with the environment,
- Use eye movements in a different manner than neck movements,

- Avoid symmetric vergence and use separate eye modules.

3.4.4 TRISH

Milios, et. al [MJT93] report a system inspired by biological vision systems. Their motivation was the need to equip a mobile platform with color binocular vision with controllable vergence. Two eyes are capable of torsion, vergence, and version movements. Torsion is the rotation of a camera about its own optical axis - this becomes useful for forcing epipolar lines to lie at a single orientation. Version is the rotation of the eyes about a vertical axis so as to maintain a constant disparity. They begin with a discussion on the geometry of convergent binocular stereo.

System Design

In typical parallel-optical axis binocular stereo arrangements, the accuracy of the extracted information is limited by the camera arrangements. It is well-known that convergent vision provides higher accuracies per unit baseline. The cost of this increased precision is the difficulty of the search for inter-ocular correspondence. The search is transformed from a simple horizontal search to a search along an epipolar line of varying slope. With the addition of cyclotorsion capabilities to the system, these non-horizontal epipolar lines can be forced to the horizontal by rotation around the optical axis.

Experimentation

TRISH (Toronto iRIs Stereo Head) contains seven degrees of freedom; however, each eye is constrained to move in concert with the other. Thus, the effective system contains only 6 DOF. They report a retinal resolution (number of detectors per unit area) of 10

Table 3.3: Primary Mechanical Capabilities of the TRISH.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	±80.0	0.003	54.0
Tilt	±35.0	0.003	54.0
Torsion	±180.0	0.003	180.0
Vergence	±35.0	0.0019	100.0
Version	±35.0	0.0019	100.0

um. which is comparable to the human retinal resolution of 7um. The baseline of the system can roam between six and thirty-two centimeters. Table 3.3 records some other interesting details.

Discussion

The authors report that TRISH has been milled and assembled. They conclude that the accuracy of the positional units is greater than the resolution of the CCD used in the imaging processes (provided that an angular position is always approached from the same direction). They also report that inertial loading makes an analytical study of the system difficult in control theory. They are currently experimenting with optimizing the head movements and using torsional degrees of freedom to solve stereo matching problems.

3.4.5 The Aalborg Head

Christensen [CM94] observes that there are two classifications of robot heads that have been constructed. The first classification, called Type I heads, are fast and accurate.

System variables are fashioned for a general purpose use. Next, he specifies a class of heads that is intended to be flexible enough for a number of different tasks in computer vision - called Type II heads. An example of a Type I head is the KTH head [PE93], while the Harvard head [FC93] is a good example of a Type II head. He further characterizes heads into a generational schema consisting of first and second generation head designs. First generation robot heads, like the Harvard head, were mainly designed from home-made components. Second generational heads like the KTH head were built with regimented design criteria as quality mechanical hardware became readily available. The Aalborg head was designed as a Type II- Second Generation head to work in a knowledge-based vision system.

System Design

In his analysis of head designs, Christensen labels degrees of freedom in active systems into three distinct categories. The first of these is called *intrinsic* camera configuration or anthropomorphically - the Eye configuration. This configuration consists of camera parameters such as focus, aperture, zoom and focal length. Next, he defines *extrinsic* camera parameters. These parameters can be likened to eye motion or position. They consist of camera tilt, cyclotorsion, and camera position with respect to the head position. Finally, he specifies head parameters such as head pan, tilt, and length of the baseline vector. His methodology in head design is to outline the importance of each of these parameters and isolate a subgroup of critical parameters that will be included in the Aalborg head. System capabilities are summarized in Table 3.4

Parameter selection for the Aalborg head was performed in a task-dependent framework. Christensen defines a set of tasks which the head will be engaged in and uses this

Table 3.4: Primary Mechanical Capabilities of the Aalborg Head.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	±90.0	0.003	200.0
Tilt	±45.0	0.003	200.0
Torsion	NA	NA	200.0
Vergence	NA	0.007	45.0
Version	NA	0.007	100.0

task list as the starting point of his analysis. His tasks were manipulation and close-range monitoring of dynamic scenes. From this he deduced that the principle depth range would be from [0.5,5] meters.

Focus speed was required to change from 0.5 to 1.5 meters in 5 seconds for tracking purposes, while focus distance was allowed to function over the common range. Aperture control was deemed crucial to the Aalborg head. Christensen provides dynamic lighting conditions and controlled depth-of-field as primary reasons for inclusion. Zoom is required to operate in the range of [5,50]mm. He uses knowledge on viewing range and deductions of desirable feature sizes at certain focal lengths to synthesize these boundaries. However, he comments that zoom is one of the features that many researches have foregone in the interest of size and weight requirements.

The ability to provide a fixation point, called vergence, which is defined by the point where the axes intersect (in a binocular system) was necessary for the applications of the Aalborg head. Here Christensen makes a distinction between second and first generation head designs. He relates that first generation heads used a symmetric vergence mechanism that provided a coupled motion of the vergence angles. He goes

on to mention that in such a system the gaze may only be changed to a position along a line perpendicular to the baseline vector (line between the focal, points also named the cyclopean axis). Thus, he states that if the goal of vision is insight into biological systems independent vergence angles is required. With respect to resolution, he states that it is desirable to have resolution of 0.01 degree in vergence angle to have depth accuracy better than one 1%. A problem in the use of non-symmetric vergence angles is that epipolar lines will have tilted with respect to the optical axis. This can be corrected by a rotation (of each camera) around the optical axis - called cyclotorsion in biological vision.

Selection of Head Parameters

Christensen points out that since the best accuracy is obtained from vergence stereo when the vergence angles are symmetric, panning of the head is required. Tilt should also be included in a head design due to similarities with the human head. He also points out that the accuracy of stereo reconstruction is directly proportional to the baseline distance; however, he acknowledges that the correspondence problem is also a direct function of baseline. Thus, he justifies the use of a controllable baseline so that initial matching could be done at short baseline distances with increased accuracy gained by iterative baseline adjustments.

The prototype design of the Aalborg head (see Figure 3.7) contains twelve degrees of freedom - six optical and six mechanical. The optical (intrinsic) parameters are capable of fully adjustable zoom, focus, and aperture. Mechanical capabilities include two vergence motors, two baseline motors, and a pan and tilt system.

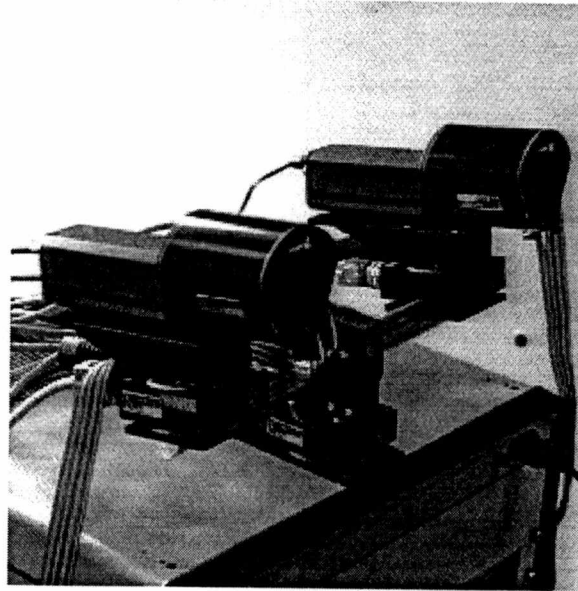


Figure 3.7: **The Aalborg Head:** A head designed with purposive constraints.

Experimentation

The first application of the Aalborg head involved the use of vergence in combination with attention control. A set of fixation points was generated from a corner detector and depth measurements were obtained at each point. Distances of the fixation points ranged from three to five meters. Christensen reports errors between 3-5% across that dynamic range. A second experiment involved mounting the platform on a translating rail. Motion stereo was then used to recover depth at an error rate of less than one percent.

Discussion

The Aalborg group has done a thorough analysis of basic head design. However, Christensen has recommended several improvements that could improve his design. These are outlined as follows:

- Use a lighter, more controllable set of lenses,
- Avoid backlash at all costs, and
- Use motors with large torque capabilities.

3.4.6 The Surrey Attentive Robot Vision System

Pretlove and Parker [PP93] employed an “integrated engineering” approach to the problem of robot head design. They begin specification with the assumption that their design will be used in a manufacturing environment, in particular it must be capable of being grasped and positioned by a standard industrial robot. They relate that these requirements will introduce additional constraints that many previous head designers have not considered. In addition to the traditional design considerations, they include *low weight* and *compact size envelope* to minimize the effect of the robot head on the dynamic performance of the manipulator.

System Design

The Surrey head has four primary degrees-of-freedom: independent vergence, focus, and aperture. However, additional degrees-of-freedom will be present in an actual application due to the presence of the robotic manipulator. As a consequence of the lack of zooming capability and the constant baseline of 200 millimeters, this design requires the positioning of the head by the manipulator to a position that puts the work area into the stereo field-of-view. The entire design weighs about 3 kilograms which the authors report as much lower than any of the previous designs. The kinematic capabilities of the Surrey head are listed in Table 3.5.

Table 3.5: Primary Mechanical Capabilities of the Surrey Head.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	±90.0	0.0	200.0
Tilt	±45.0	0.003	200.0
Torsion	NA	NA	200.0
Vergence	60	0.07	45.0
Version	NA	0.007	100.0

Experimentation

Experimentation on the Surrey head consisted of tracking an object along a straight line of variable depth with respect to the robot's end-effector. An implementation tracking a white object on a black background with cycle times of approximately 70 ms was reported. Future experimentation will revolve around the implementation of a robotic docking procedure that will automatically mount the stereo head on the end-effector.

Discussion

The Surrey head represents a promising approach to head design. The authors have stated their design goals in a manner that is directly related to their primary task. Their design makes implicit use of an industrial manipulator as part of the overall kinematic package. This has the advantages of reducing the complexity of the head design while providing capabilities similar to a kinematically more complex head. The primary disadvantages of the Surrey head include a fixed baseline and the lack of zooming capabilities. The integrated design is shown in Figure 3.8.

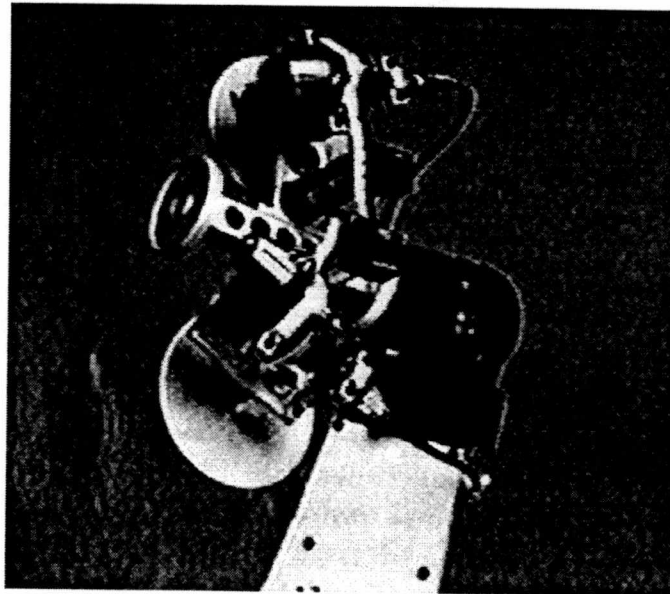


Figure 3.8: **The Surrey Attentive Robot Vision System:** A head designed with and integrated engineering specification.

3.4.7 The University of Pennsylvania Head

The University of Pennsylvania [KFS88, Kro89] reports the development of a robotic head with two more translational and one less rotational degrees-of-freedom than the human neck. Table 3.6 outlines a few of the basic capabilities of its mechanical system. Construction of the system was motivated by the need to acquire quality images across a large variety of scenes. The authors intend to use the optical and mechanical degrees-of-freedom to acquire adequate views of objects which may suffer from inadequate lighting and partial or complete occlusion.

System Design

Five degrees-of-freedom are present in the Pennsylvania design. Four of these are contained within the camera platform, while the remaining DOF is provided by an an-

Table 3.6: Primary Mechanical Capabilities of the Pennsylvania Head.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	63.8	0.0	NA
Tilt	49.6	0.003	NA
Vergence	5.5	0.08	NA
X Trans	36.8 cm.	NA	NA
Y Trans	25.4 cm.	NA	NA

tisymmetric vergence mechanism. The platform allows translation in both principal directions of the image plane. Along with these, the more traditional pan and tilt rotations of the stereo system are provided. Motorized lenses occupy an integral part of the design allowing 6 times zoom capability. The most unusual design feature of the Pennsylvania system is the addition of a computer controlled lighting system. Ten lamps whose intensity can be individual set using duty cycle control are used to image objects with a large range of reflectance properties.

Experimentation

Krotkov uses the Pennsylvania system as an aid in solving the *spatial layout* problem. In particular, he has developed modules for depth from focus and vergent stereo. In addition to the work done on each of these modalities, a cooperative ranging module that autonomously computes ranges by cross-checking and statistical testing has been implemented. His experimentation is believed to be the first “dynamic, active, conjunctive” operation of two visual sensing modalities. Cooperative ranging results showed dramatic increases in accuracy of the ranging results obtained by any one of the single

range measurements.

Discussion

The Pennsylvania system was one of the first attempts at robotic head design. The author admits, "it was built at a low cost, largely from surplus equipment and spare parts". As a result, most of the components lack sufficient mechanical precision for applications in photogrammetric stereo; however, the general ideas behind this system have had a tremendous impact on active vision platform development. In addition the experimentation done by the Pennsylvania group is among the most extensive of the reviewed groups. A complete reference to this Krotkov's work can be found in [Kro87].

3.4.8 BETH

The Computer Vision and Robotics Research Laboratory at the University of Tennessee, Knoxville has assembled an integrated sensory-motor platform capable of independent vergence and accommodation [MTL93a]. The platform, called *Binocular Eyes for Tennessee Head* (BETH) maintains four degrees of mechanical freedom on independent pan and tilt units and six degrees of optical freedom obtained from motorized lenses. Table 3.7 shows the primary mechanical capabilities of the BETH system. BETH was designed to not only to provide a flexible image acquisition platform, but also to be used in a dynamic environment where both the robot and objects in the environment are possibly moving. Marapane claims that accommodation, fixation, and vergence are critical for processing information in such an environment, and in many cases their presence will simplify many perceptual tasks [MT94a].

Table 3.7: Primary Mechanical Capabilities of BETH.

Axis	Range	Resolution (°)	Speed (°/s)
Pan	±159.0	0.003	200.0
Tilt	±35.0	0.003	200.0
Vergence	NA	0.007	45.0
Version	NA	0.007	100.0

System Design

BETH's mechanical systems consist of two calibrated, ASCII controlled pan and tilt units. These units provide movements similar to human lateral-medial and superior-inferior rotation. Control is limited to the issue of 9600 baud serial commands. The stereo system was constructed by mounting both pan/tilt units on a section of optical bread board. Baseline distances can varied manually but must reside on one inch increments due to the discrete nature of the mounting plate. Optical capabilities are provided by two programmable variable zoom lenses. Each lens is mounted onto a pan/tilt unit, thus giving the independent vergence capability that many advanced applications may require. Again, the lens control is limited to 9600 baud serial connections. Fully assembled, BETH offers ten degrees-of-freedom to the Tennessee group. BETH is shown in Figure 3.9.

Experimentation

The most singularly unique feature of the work done on BETH is the extensive use of simulation in the development of active algorithms. Marapane [MTL93b] has developed both real and virtual BETH systems. In this manner, algorithms are developed and

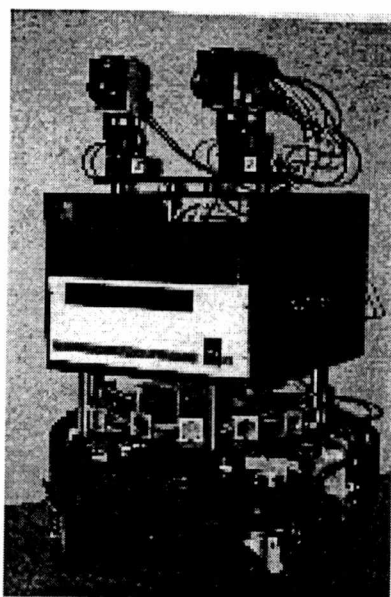


Figure 3.9: **Binocular Eyes for Tennessee Head (BETH):** A multi-purpose head for use in tracking and robot conveying.

debugged on the virtual BETH. After this rough testing is completed, the algorithms are ported to the real system for fine tuning and verification.

Testing of BETH has revolved around the concept of active tracking. Marapane has developed a robot conveying [MT94b] application that tracks an infra-red led on the back of a teleoperated robot. His algorithm uses the monocular capabilities of BETH integrated with a single ultrasonic transducer to return approximate depth estimates. Horstmann, et. al [JHTRT95] again used the monocular facilities on BETH to provide a predictive tracker using an implementation of the Kalman filter in real-time. The tracker used a sum-of-squared difference correlator integrated with the predictor to accurately track pixel blocks from one frame to another.

Discussion

BETH is a versatile head design that has been tested in several different applications. The most unique aspect of the UTK system is the existence of a complete graphical simulation of the hardware system. In fact, most development is done on the simulated BETH to insure algorithmic integrity before moving to the real device. Also the UTK group has done a very thorough job of testing the applicability of BETH to a variety of active vision problems. Extensive real-time implementations, particularly trackers, have been applied to research domain problems. BETH has also been employed in static configurations as a setup for parallel-axis stereo [TT95a] and inverse perspective mapping [Ric96] applications.

3.4.9 PENNEYES

In addition to the work done at Pennsylvania by Krotkov [Kro89] in the late eighties, Madden and von Seelen have designed a robot head in conjunction with the Universities of Rochester, Maryland, Massachusetts, and Transitions Research Corporation (TRC). The PENNEYES system is a two axis version of the BiSight head sold by TRC. Figure 3.10 displays the two-axis head as used in a stereo tracking application. Kinematic capabilities are augmented through the use of a 6 axis PUMA 560 robot. The major goal of this joint project was to benefit research through the development and sharing of a common vision platform.

System Design

The PENNEYES version of the BiSight Head has an independent vergence mechanism provided to each of the camera mounts. Imaging capabilities are provided through

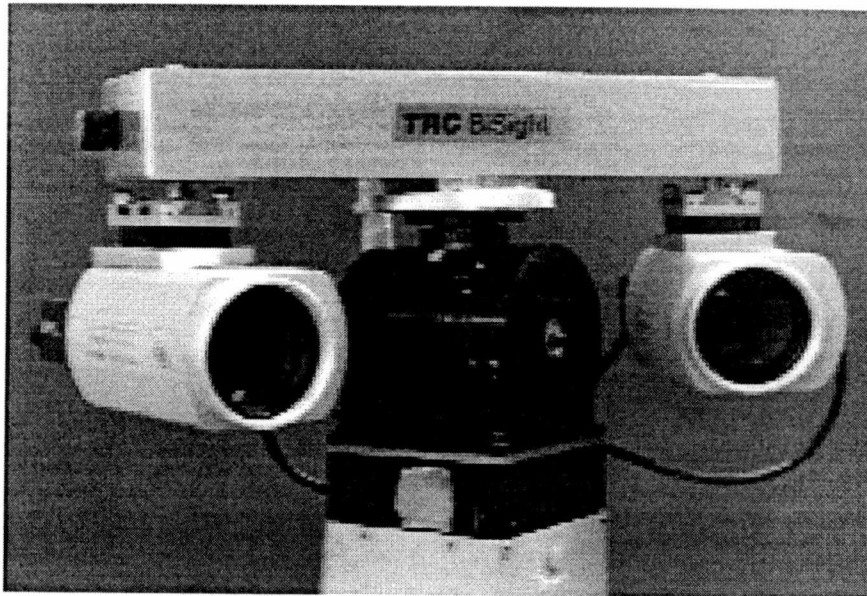


Figure 3.10: **The PENN EYES Robotic Head:** A commercially available head design.

standard CCD cameras and are supplemented by the addition of motorized zoom lens with the potential for ten fold magnification. In addition, provision has been made to mount external lenses to the assembly for smaller viewing ranges. The system is also equipped with an eight axis motion controller for simultaneous command of up to four optical and four mechanical DOFs.

Experimentation

Initial experimentation with PENNEYES involved the implementation of a tracker of single object, high contrast scenes. Using a simple thresholding and centroid calculation, it was possible to achieve field rate tracking under these relaxed conditions. As a next generation of experiments, Madden has chosen the problem domain of arbitrary targets against arbitrary backgrounds. Currently, they have implemented a field rate

computation of visual error. This system is designed to provided redundant tracking of an object moving in three dimensions.

Discussion

Kinematically, the PENNEYES system is the most limited of the presented platforms (when decoupled from the PUMA robot). Its main advantage is that it was derived from a set of commercially available robot heads that are currently being developed. The availability of inexpensive, quality active vision hardware should have the affect of stimulating new research into the process and algorithms of active stereo vision. The experimentation done by the Pennsylvania group on PENNEYES involved pseudo-real time implementations of some standard tracking algorithms.

3.5 Discussion

A basic review of the "state-of-the-art" in active platform design has been completed. Some designs, i.e. KTH head, TRISH, and the Harvard head where designed to be anthropomorphic in order to implement the mechanisms of low-level human perception on a kinematically similar platform. These systems where not design with a specific task goal in mind and therefore fall outside the realm of purposive systems. The Aalborg head and the Surrey system where implemented in a task-domain framework, albeit from two distinctly different perspectives. The Aalborg head was designed with a generic set of tasks in mind. These tasks, manipulation and close-range tracking, were used to heuristically guide his design process. While the Surrey system used strict mechanical constraints, such as weight limit and a minimum volume envelope, it did constrain the

stereo design variables in any way as a function of task. The Pennsylvania system and BETH were both design to solve generic vision problems. Krotkov uses his system to study the spatial layout problem, while the UTK team used the BETH system to study a wide range of general vision problems in an active framework. Finally, the PENNEYEs design is an off the shelf system, designed for general purpose use. It is low-cost and efficient; however, for many tasks it may suffer from its highly generic design.

What is needed in purposive vision is a design guideline for implementing active systems. Designs must be framed in a task. Only then can design work begin to produce systems with stability and robustness in the face of dynamic environments. In the next sections purposive design is explored in four specific contexts. First, parallel-axis stereo systems are studied followed by convergent-axis stereo platforms. The fruit of this labor is a procedure for designing robot heads as a function of task. Given that the procedure is followed, the design will be optimal under derived constraints. Next, the calibration problem is studied in detail, and a heuristic for maintaining the integrity of data is presented. The third section of analysis looks at a generic class of representations to perform basic functions such as template matching and region formation. Finally, an algorithm for the design of purposive systems in a Bayesian framework is highlighted. Using this procedure, it is possible to produce robust model-based vision using a connected network of simple estimators and numerical optimization tools. Taken as a whole, this work can be viewed as a guidebook for the production of purposive vision systems. These systems will provide a higher degree of reliability than those designed for a generic task and then used to implement specific functionality.

CHAPTER 4

Optimal Geometries for Parallel-Axis Stereo

4.1 Introduction

Suppose that you are given the task of selecting a particular geometry for a parallel-axis stereo vision system. If this task is approached from a simplified perspective of stereo vision and using only a layman's knowledge, you could affix two conventional cameras onto a support structure and roughly align their optical axes. Then, dust off a few old lenses that you were lucky enough to find in a lab drawer and mount them on the cameras. Then, call your boss and amaze him/her with your industriousness. Naturally, he is impressed with your ability and selects your arrangement to give range measurements for the company's milestone robotic tank firing demonstration. After narrowly escaping the demonstration, you and your boss are fired.

Or, the task could be approached from an traditional academic perspective which can be strikingly similar to the previous analogy. Being an industrious graduate student and interested in building the most accurate stereo vision system possible, I spend countless hours researching and quantifying sub-pixel precision localization algorithms and a bullet-proof solution to the stereo-matching problem. Now since no self-respecting graduate student would turn in a two-hundred page report void of experimental data, I tape two cameras to my hockey stick, make a few measurements, and find that my setup has the accuracy that my advisor wanted to see. Of course, he misses the footnote

on page 146 where I have described that the range measurements were obtained using object distances between 3 and 4 inches away from the cameras.

Aside from being mildly humorous, these stories serve to illustrate two fundamental problems that can be easily overlooked in the design of ranging systems that use stereo vision. While both systems certainly satisfy the minimal definition of a stereo vision system, they will most likely fail when applied to a real problem. In the first case, the problem was not properly specified in the design phase; consequently, the project was doomed to failure. In the second, some constraints were specified; however, they were not formulated in any meaningful context.

Thus we can consider a system adequate if it was designed using sufficient constraints obtained from a proper specification of the task. Christensen concurs saying, "to provide a framework for the specification of parameters a specific scenario is considered" [CM94]. Relating this to stereo vision, how can the proper constraints be obtained from a definition of a task? Clearly, the general task demands that precise three-dimensional range data be recovered [MT95]. Coupling this with a specific application may require that other variables such as the binocular field-of-view, minimum object standoff, and maximum allowable error limits are balanced in an appropriate fashion. This work addresses both issues. First, a system of flexible constraints is derived that addresses the needs of a wide variety of stereo-ranging applications. Using these constraints and a simple model of a stereo system, a closed-form solution is derived that offers the optimal geometry for parallel-axis stereo configurations under the assumed conditions. An algorithm for Parallel-Axis Stereo Synthesis (**PASS**) is outlined. This algorithm can be used in at least two important ways:

- Off-line generation of static parallel-axis stereo geometries that are optimized for

a particular task.

- On-line generation of active parallel-axis stereo configurations that are optimal for the current task.

This paper addresses only the first of these possible applications. The text concludes with both simulated and experimental verification for a sample robotic task. As a first step, the geometrical structure of a simple stereo model is presented upon which all derivations and experimental results are presented.

4.2 A Simple Model for Parallel-Axis Stereo Geometry

In [FGL87], a simple, illustrative first-order model of the geometrical relationships in a parallel-axis stereo configuration is presented. Figure 4.1 represents the basic geometrical structure of this configuration. In the figure, the quantities f , ρ and b , which represent focal-length, physical dimension of the image plane, and baseline, respectively, are the physical variables that completely specify the geometry. (Note: ρ is assumed to be a constant since most "off-the-shelf" cameras have similar image plane dimensions.) ΔZ , ΔF , and $Z_{\min}$ represent the purposive constraints that will be fully developed in the next section.

Previously, it was stated that the general goal of ranging systems is to recover the most accurate range data that is possible. Horn [Hor89] states that the accuracy of the recovered range measurements increases with increased baseline and focal length. Thus, for fixed object distances we can increase the accuracy by increasing the product of baseline and focal-length. Unfortunately, larger baseline distances and focal-lengths cause proportional increases in the minimum distance which objects must be from the

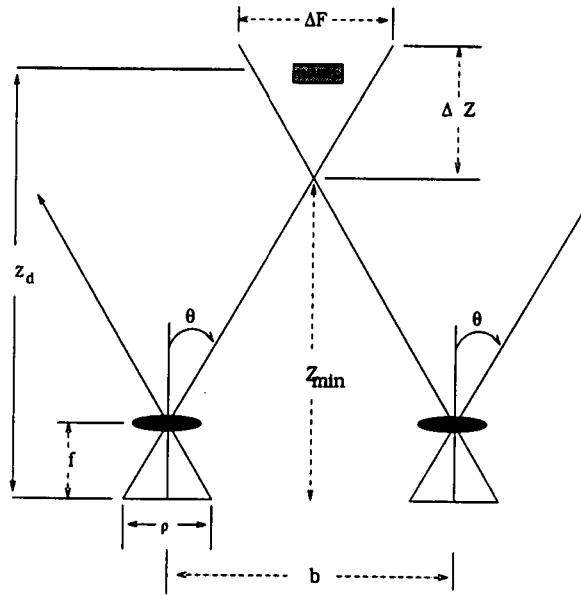


Figure 4.1: **Basic Geometry:** Stereo vision system.

array before they are imaged in both cameras. The basic relationship that represents these tradeoffs is

$$b\Delta Z = Z_{min}\Delta F. \quad (4.1)$$

Here, b is the baseline distance. ΔZ is a component included to limit the field-of-view to the dynamic range of the robotic workspace. Z_{min} represents the start of the workspace volume and the distance to the point where stereopsis begins. ΔF is a measure of the sensitivity of the configuration. If it is small the stereo disparities that are generated will produce highly accurate range data at the expense of the area of stereo sensitivity. Also notice that the focal length is represented implicitly.

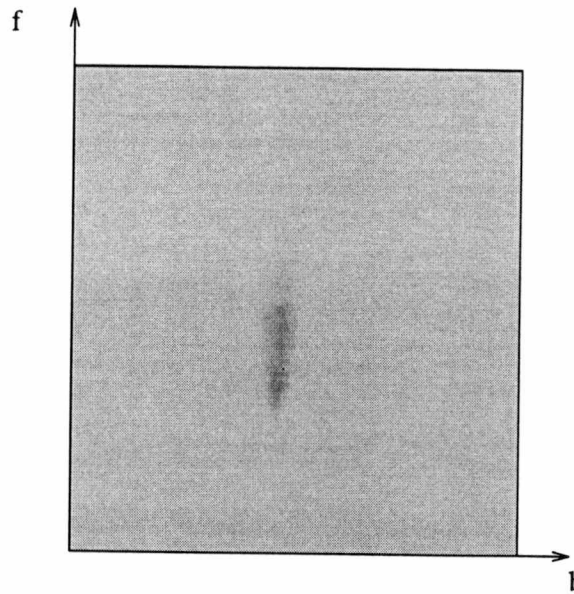


Figure 4.2: **Geometrical Design Space:** Without meaningful constraints, the design space is infinite. (No stable design points can be located)

4.3 Isolating Task-Based Constraints

As previously outlined, the parallel-axis stereo geometry can be completely outlined by specifying two variables: baseline and focal length. The abstracted design space for the problem is represented in Figure 4.2. As shown, there are an infinite number of possible combinations of baseline and focal-length. The basic task in design is to reduce this space to a single point which will represent a specific parallel-axis configuration. Without task-based constraints, the problem is clearly intractable, and its solution is relegated to intuition. The application of task-based constraints will serve to limit the design space to a more manageable form and should improve the quality of the resultant stereo vision machine.

The first set of constraints that should be applied are done in a typical design process.

They are outlined here for clarity. Although the theoretical design space is infinite, a real system cannot possibly achieve such a wide array of configurations. It is limited in both baseline and focal-length. Baseline is constrained by the amount of space available for the stereo system, while focal-length is usually restricted by the availability of “off-the-shelf” optical components. Regardless of the physical quantification of the limits, (which will vary widely from application to application) it is clear that each parameter can be constrained to lie on an interval (see Equation 4.2).

$$\left\{ \begin{array}{l} b_{\min} \leq b \leq b_{\max} \\ f_{\min} \leq f \leq f_{\max} \end{array} \right\} \quad (4.2)$$

This results in a reduction of the design space as shown in Figure 4.3. This can be considered a simple mapping of the problem into a set of configurations that can be physically realized in a given scenario. However, the number of configurations is still unbounded and must be restricted by further constraints.

The next constraint that should be included is obtained directly from the geometry of the setup in Figure 4.1. For a given baseline-focal length pair, a minimum standoff to the area of stereopsis is implied. Thus, the addition of a interval representing the limits of this standoff further constrains the design space. This constraint also has a direct relationship with the application. For example, most robotic manipulators have a limit on their working volumes. The standoff constraint can be related to the distance from the stereo system to the centroid of this volume. In this way it can be directly obtained from the kinematics of the robotic manipulator. Figure 4.4 displays the design space after application of the standoff interval constraint. Equation 4.3 formalizes this constraint in the representation model, where ρ is the quantity representing the physical

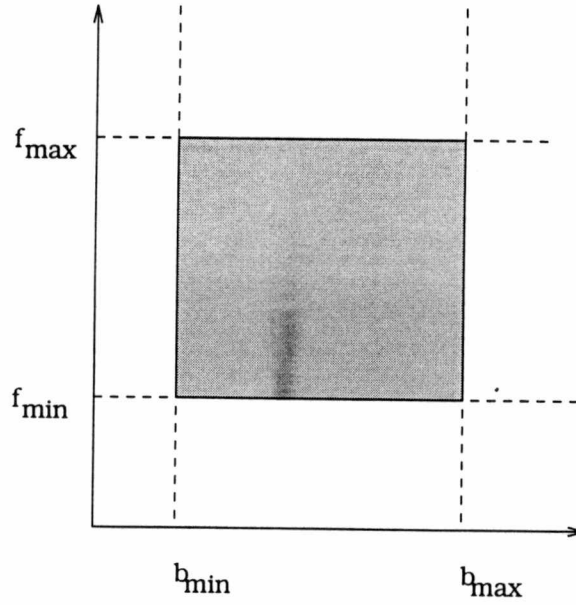


Figure 4.3: **Geometrical Design Space:** The addition of simple physical constraints offers a more manageable design space; however, the search space has been restricted in a trivial context

dimension of the imaging device. Each point on the standoff boundary line in Figure 4.4 has the same baseline-focal length product. Since accuracy is directly related to this product, each of these points will produce the same quality of results in the model at the point where stereopsis begins; however, as will be shown, this line can be reduced to a single point with the addition of some working volume constraints. Not only do these constraints represent physical quantities that are readily obtained for the task, their application can be shown to produce a maximum in the baseline-focal length product.

$$Z_{\min} = \frac{bf}{\rho} \leq Z \quad (4.3)$$

In Figure 4.1, the area of stereopsis is represented (in two-dimensions for convenience) by a simple triangle that can be assumed to be isocles if the focal-lengths of the stereo cameras can be assumed to be equivalent. While this geometry is fixed by the parallel-

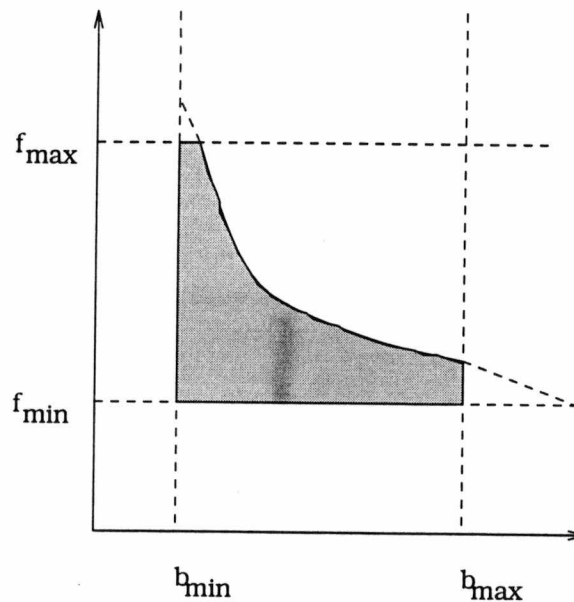


Figure 4.4: **Geometrical Design Space:** Here the design space has been constrained by a task oriented constraint (standoff: equation);however, each point along the curve can be considered optimal

axis configuration, it may not be suitable as a representation of workspace geometry. These work volumes can take on an arbitrary shape depending on the application, available hardware, and designer specification. It is intended, however, to assume primitive workspace geometries that will be of some practical use, and try to offer the “best-fit” of the stereo triangle to that particular shape. Naturally, this fitting process should be directly motivated by the desire for the most accurate range data that can be obtained from the input constraints.

4.3.1 A Rectangular Work Space Constraint

A common geometry that can adequately represent many applications is the square, or more generally the rectangle. Examples of actual systems that have rectangular geome-

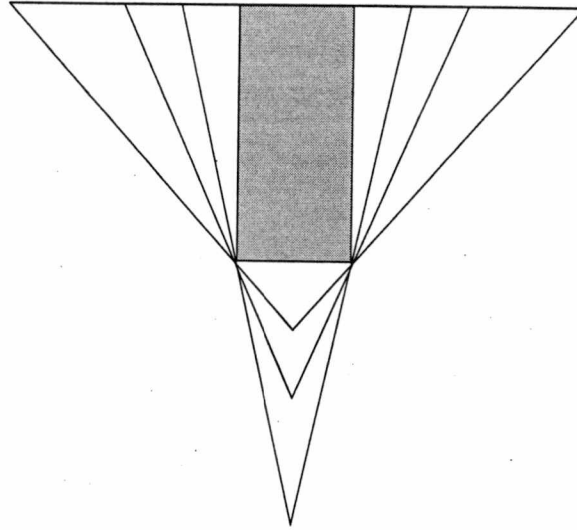


Figure 4.5: **Circumscribing an Isoceles Triangle onto a Rectangle:** An infinite number of solutions exist. However, Area and Perimeter have a distinct minima.

tries are cartesian robots and many industrial machine vision systems. Typically, the geometry of the rectangle can be determined directly from the desired task. Therefore, if given a rectangular work area, the problem becomes fitting the best isosceles triangle to that area with respect to the accuracy of the stereo system. Unfortunately, as shown in Figure 4.5, the number of isosceles triangles that can be circumscribed onto a rectangle is infinite. However, closer examination reveals that the area and perimeter become infinite as the base of the triangle approaches the width of the rectangle or as the height of the triangle approaches the height of the rectangle. Since the area or perimeter equations are differentiable, a distinct minimum should exist.

Assuming that the length-to-width ratio of the rectangle is α , the relationship of the base of the triangle to its height is

$$\Delta F = \frac{\Delta Z w}{\Delta Z - \alpha w}. \quad (4.4)$$

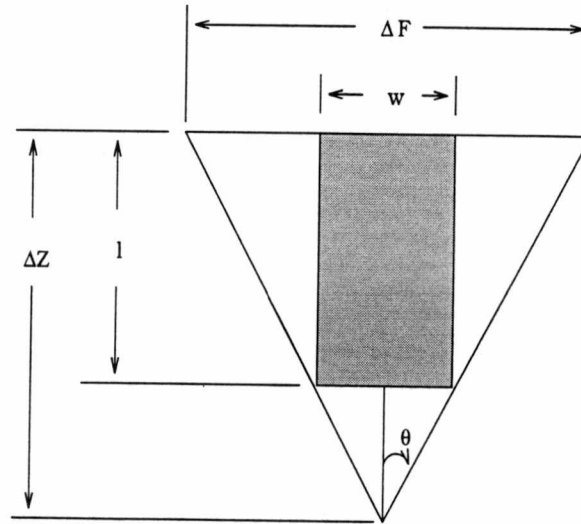


Figure 4.6: **Circumscription of an Isocles Triangle onto a Rectangle:** The geometrical relationships superimposed with the task-based constraints give a direct solution to the stereo geometry problem.

Substitution of Equation 4.4 into the formula for the area of a triangle gives

$$A_t = \frac{\Delta Z^2 w}{2(\Delta Z - \alpha w)} \quad (4.5)$$

In Equation 4.5, w and α are constants with respect to the area of the triangle; therefore, the minima can be obtained through direct means. Differentiating Equation 4.5 and equating it to zero produces

$$\left\{ \begin{array}{l} \Delta Z = 2\alpha w = 2l \\ \Delta F = 2\frac{l}{\alpha} = 2w \end{array} \right\}. \quad (4.6)$$

Figure 4.6 depicts the geometrical relationships that were derived in Equation 4.6. As a consequence of the assumption of equivalent focal lengths in each stereo pair, the ratio of ΔZ -to- ΔF equals l -to- w (in the case of a square, where $l = w$, the ratio is 1).

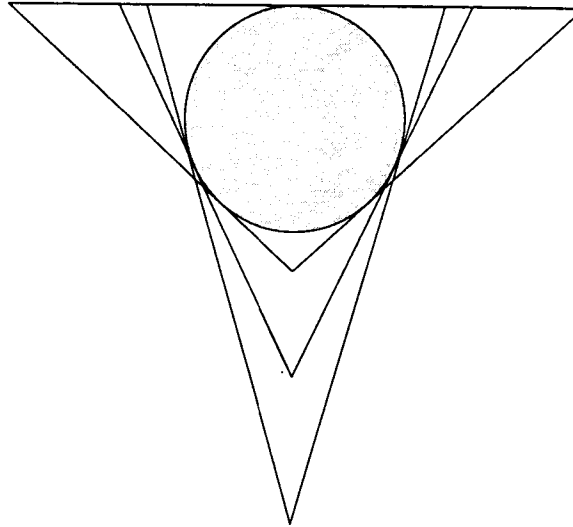


Figure 4.7: **Circumscription of an Isosceles Triangle onto a Circle:** An infinite number of solutions exist. However, area and perimeter have a distinct minima.

4.3.2 A Circular Work Space Constraint

Another geometry that is common in many applications is the circle. Many robots have spherical work volumes and many ranging applications have a central region of interest and a periphery that can be approximated by a circle or sphere. Figure 4.7 displays the same preponderance of solutions for the range triangle as in the case of the square. Again, the area and the perimeter can be uniquely minimized. Equation 4.7 represents the radius of the maximum circle that can be inscribed in an isosceles triangle with sides ΔF and a . (Note: s represents the semi-perimeter).

$$r = \frac{(s-a)\sqrt{s(s-\Delta F)}}{s} \quad (4.7)$$

$$s = a + \frac{\Delta F}{2}$$

Solving for a in Equation 4.7 results in

$$a = \frac{\Delta F}{2.0} \left\{ \frac{\Delta F^2 + 4r^2}{\Delta F^2 - 4r^2} \right\}. \quad (4.8)$$

Using the pythagorean theorem to solve for a in terms of the base and height of the triangle gives

$$a = \sqrt{\Delta Z^2 + \frac{\Delta F^2}{4}}. \quad (4.9)$$

Substitution of Equation 4.7 and Equation 4.9 into the area formula for a triangle results in

$$A_c = \frac{\Delta F^3 r}{\Delta F^2 - 4r^2}. \quad (4.10)$$

Similarly to the previous case, the radius r in a given problem, can be considered to be constant. Solving for the minimum area results in

$$\left\{ \begin{array}{l} \Delta F = 2\sqrt{3}r = a \\ \Delta Z = 3r \end{array} \right\}, \quad (4.11)$$

where the optimal case is the equilateral triangle. Figure 4.8 depicts the geometrical structure of the solution found from Equation set 4.11. An interesting artifact of the choice of circular geometry is that the focal-length is independent of the constraint variables and remains constant at a normalized value of $\frac{\sqrt{3}}{2}$. (Note: normalized focal-length is defined as $\frac{f}{p}$.) This is because the resulting triangle is equilateral.

4.4 The Parallel-Axis Stereo Synthesis (PASS) Algorithm

Now that a closed-form, optimal solution has been derived, an algorithm for generating stereo geometries given the input constraints is presented. Figure 4.4 outlines its

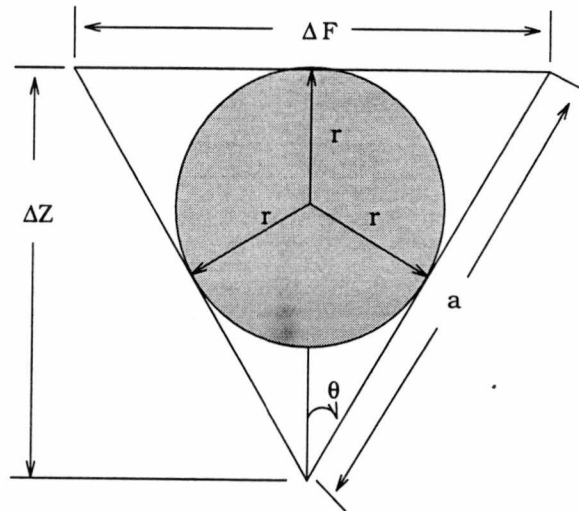


Figure 4.8: **Circumscription of an Isosceles Triangle onto a Circle:** The geometrical relationships superimposed with the task-based constraints give a direct solution to the stereo geometry.

structure. Success of the algorithm is linked to a proper match of input constraints to the task. In the PASS algorithm only two major constraints are considered: work space geometry and the average viewing distance. It is critical that they be properly specified in any particular application. As shown in Figure 4.4, the recommended first step is a careful selection of the viewing geometry.

4.4.1 Specification

Essentially, this step involves the choice of circular or rectangular geometry and specification of its representative parameters. Selection of the desired geometry is primarily a task of finding the closest match between task space and the available primitives, in this case a rectangle, square, or circle. Once the basic shape of the workspace has been chosen, the expanse of the work space must be determined. An important “rule-of-thumb” is to keep the workspace as small as possible for the particular task. In doing

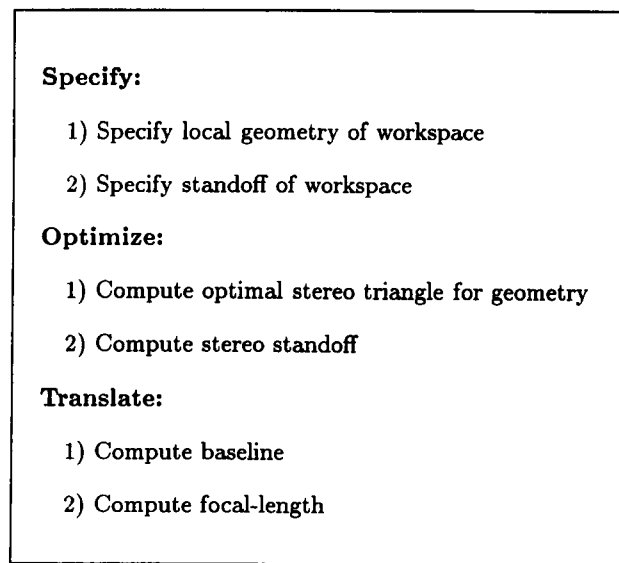


Figure 4.9: **Parallel-Axis Stereo Synthesis (PASS) Algorithm:** Systemic constraints are mapped into unique stereo geometries using a minimum area difference solution.

this, a premium is placed on the accuracy of the resultant range data. However, as in all vision algorithms there is an identification/localization tradeoff where precise results are obtained at the expense of the probability of locating objects.

Figure 4.10 offers one interpretation of this tradeoff. In the figure, three different geometries are depicted and overlayed onto a particular scene. The smallest stereo triangle (which is also the most precise) views only a small proportion of the scene. In contrast, the larger triangles are capable of viewing larger cross-sections of the scene at the expense of precision in the recovered range data. This uncertainty is expressed analytically in Equations 4.14 and 4.15 where baseline is a function of the ratio ΔF -to- ΔZ and focal-length is expressed in terms of the inverse ratio.

One way to derive these parameters is by looking at the geometry of typical objects that the system will be viewing and then setting the geometrical parameters to some

multiple of those. For instance, if the scene is to contain objects of less than 4 inches in maximal span in any direction , then one can select a circle of radius 5 inches to provide highly accurate numbers.

4.4.2 Optimization

This stage of the algorithm produces the optimal stereo triangle and the standoff using input constraints and the derivations from Section 3. First, the ΔF and ΔZ parameters are computed for the stereo triangle. In the case of a rectangle or a square, Equation set 4.6 is employed, or Equation set 4.11 is used if the geometry is circular. Finally, the average distance to the work space, which is defined as the distance from the stereo reference to the centroid of the work space (Z_d), can be related to the offset at which binocularity begins, $Z_{\min}$. For the rectangular or the square cases, this conversion is

$$Z_{\min} = Z_d - \frac{3l}{2}, \quad (4.12)$$

and for the circle,

$$Z_{\min} = Z_d - 2r. \quad (4.13)$$

These conversions force the entire work space to be within the area of stereo sensitivity. Now, all that remains is the calculation of the physical stereo parameters.

4.4.3 Translation

Parallel-axis stereo geometry is completely specified by a baseline and a focal length. The final stage of the PASS algorithm translates the optimal triangle parameters, ΔF and ΔZ , along with $Z_{\min}$ into these. Baseline is computed directly from Equation 4.1:

$$b = \frac{Z_{\min} \Delta F}{\Delta Z}. \quad (4.14)$$

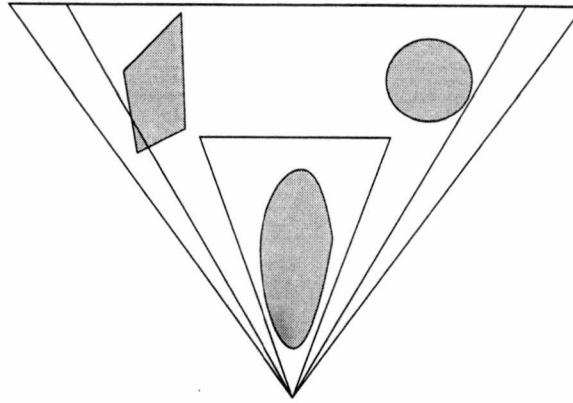


Figure 4.10: **Identification/Localization Tradeoff:** As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.

Focal length is found using

$$f = \frac{\rho \Delta Z}{\Delta F}, \quad (4.15)$$

where ρ is a constant describing the size of the image plane.

4.5 A Geometric Description of the Optimality

The optimality of this algorithm can be described by a simple geometric argument. In this framework, the constraints are decoupled in the sense that one regulates baseline and one regulates focal-length. In particular, the geometry of the work space will completely specify the normalized focal-length for the system. This is a natural extension of the assumption of equivalent focal-lengths in each camera. The algorithm simply chooses the largest focal length that will preserve the circumscription constraints. Baseline is chosen as a function of the minimum distance to the active stereo area and a simple multiplication by the base-to-height ratio of its triangle. By definition, it is the

only baseline point that will satisfy all constraints. In two-dimensional terms, the optimization occurs in a fit of the working geometry to the triangular geometry of the stereo model. However, this fitting translates directly into a maximum focal-length and subsequently a fixed and maximum baseline. Thus, in terms of our area-based fit the product of the baseline and the focal-length is a maximum. Alvertos and Gonzalez [ABG89] confirm this argument when they relate directly the “common viewing area” to the error in estimating the three-dimensional coordinates of objects within that area.

4.6 Static Parallel-Axis Configuration

In this section, a hypothetical problem that examines the results given in the derivations of the previous sections is introduced. The problem requires the specification of a stereo configuration for a robotic work cell. The standoff, Z_d , is assumed to be fixed to a reasonable distance and the remaining variables are examined as the dimensions of the work space are varied. In this way, a typical application in robotics is analyzed under three-different workspace constraints (circular, square, and rectangular). The results are outlined in tabular form for convenience. Note: All problems assume that the standoff distance (the distance from the stereo system to the centroid of the workspace) is fixed to 4 m for safety reasons. It could be set to any positive value.

Using a circular workspace constraint where the radius, r , is varied from 0.1 to 2.0 meters. Geometrical data for the minimum standoff, $Z_{\min}$, and baseline, b , is displayed in Table 4.1. The findings for the use of circular constraints are outlined:

- Regardless of the constraints, the normalized focal-length is a constant, $\frac{\sqrt{3}}{2}$.
- Since the focal-length is constant, the most accuracy is obtained by using the

Table 4.1: **Parallel Axis Geometry with Circular Work Space:** The standoff is held constant, and the normalized focal length is constant as a consequence of the equilateral geometry of the triangle.

r	0.1	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$Z_{\min}$	3.8	3.6	3.2	2.8	2.4	2.0	1.6	1.2	0.8	0.4	NA
b	4.37	4.14	3.56	3.23	2.76	2.3	1.84	1.38	0.92	0.46	NA

smallest possible working area.

- Increases in the radius result in proportional decreases in the baseline distance.

Next, a square area is tested in a similar manner. Table 4.2 contains the findings. Again the normalized focal-length is constant function of the length, l , of a side of the square, independent of the geometrical constraints. Its normalized value is a constant, 1. Also, the values of baseline and minimum standoff, $Z_{\min}$ are forced to be equal. The results are highlighted as:

Table 4.2: **Parallel Axis Geometry with Square Work Space:** The standoff is held constant, and the normalized focal length is constant as a consequence of the isocetes geometry of the triangle.

l	0.1	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$Z_{\min}$	3.85	3.7	3.4	3.1	2.8	2.5	2.2	1.9	1.3	1.0	NA
b	3.85	3.7	3.4	3.1	2.8	2.5	2.2	1.9	1.3	1.0	NA

- Regardless of the constraints, the normalized focal-length is a constant, 1.
- Since the focal-length is constant, the most accuracy is obtained by using the smallest possible working square.

- Increases in side length result in equivalent decreases in baseline distance.

The most interesting results were found in the case of a rectangle. Tables 4.3 and 4.4 contain the configurations generated from using four different width-to-length ratios, where length was allowed to vary from 0.2 to 1.0 meter.

Table 4.3: **Parallel Axis Geometry with Rectangular Work Space I:** The stand-off is held constant at and the normalized focal length is constant for each α as a consequence of the isocles geometry of the triangle.

l	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0
w	0.5	1.0	2.0	2.5	0.25	0.5	0.44	0.55
$Z_{\min}$	3.7	3.4	2.8	2.5	3.7	3.4	2.8	2.5
b	1.48	1.36	1.12	1.0	2.96	2.72	2.24	2.0
$\frac{f}{\rho}$	2.5				1.25			

Table 4.4: **Parallel Axis Geometry with Rectangular Work Space II:** The stand-off is held constant at and the normalized focal length is constant for each α as a consequence of the isocles geometry of the triangle.

l	0.2	0.4	0.8	1.0	0.2	0.4	0.8	1.0
w	0.14	0.28	0.57	0.71	0.11	0.22	0.44	0.55
$Z_{\min}$	3.7	3.4	2.8	2.5	3.7	3.4	2.8	2.5
b	5.18	4.76	3.92	3.5	6.66	6.12	5.04	4.5
$\frac{f}{\rho}$	0.714				0.555			

- For each rectangular specification, the normalized focal-length is a constant.

- Since the focal-length is constant, the most accuracy is obtained by using the smallest possible rectangular area.
- Increases in length result in equivalent increases in ΔZ .
- Increases in width result in proportional decreases in baseline separation.

4.7 Conclusions

A framework for the generation of parallel-axis stereo geometries was presented. Two primary task-based constraints were used to produce a closed-form solution to the stereo-design problem. This solution used a minimum area match of the work space geometry with the inherent triangular geometry of an equivalent focal-length parallel axis stereo system. From this solution, an algorithm for Parallel-Axis Stereo Synthesis (PASS) was outlined. Using PASS, a designer can translate work space and standoff constraints into an optimal stereo design.

CHAPTER 5

Optimal Geomtries for Convergent Stereo

5.1 A Model for Convergent-Axis Stereo Geometry

Figure 5.1 represents, in two-dimensions, the fundamental geometry of convergent stereo. The concept is similar to that of parallel-axis stereo [Hor89, FGL87] with the exception of allowing a rotation of the each camera about the vertical axis of the lens center [ABG89, Kro89, KFS88]. In the figure, two cameras are *fixated* on point **P**. The region of stereo sensitivity is formed by the polygon **ABCD** which is completely defined by the set of stereo variables $S = \{\theta_l, \phi_l, \theta_r, \phi_r, b\}$ assuming that the image plane dimension ρ is constant. The angles θ_l and θ_r are called the *convergence angles* and they represent the rotation of the cameras about the vertical axis through the lens center. Focal length in the left and right cameras is expressed through the variables ϕ_l and ϕ_r , respectively. Finally, baseline, or the inter-camera separation, is denoted by b .

The major goal of this section is to develop a model for each of the critical points which are defined as the set of points **A**, **B**, **C**, **D**, **P** from the figure. Later, this information will be used to derive a model for the area of the stereo polygon **ABCD** that can be used in conjunction with some task-domain constraints to define an optimal configuration of the variable set S . Equation set 5.1 contains the equations of the principle optical rays of the left camera system in slope-intercept form with respect to the left camera coordinate system. (For convenience the subscripts of all left camera

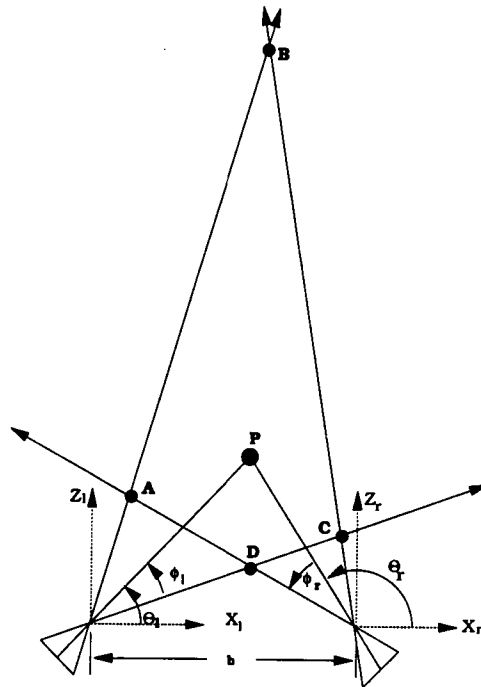


Figure 5.1: **Convergent Stereo Geometry:** As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.

coordinate lines are even, while the subscripts of the right camera coordinate lines are odd.)

$$l_0 \doteq z_1 = \tan(\theta_1)x_1$$

$$l_2 \doteq z_1 = \tan(\theta_1 - \phi_1)x_1 = mx_1 \quad (5.1)$$

$$l_4 \doteq z_1 = \tan(\theta_1 + \phi_1)x_1 = nx_1$$

Similarly, equation set 5.2 contains equations for the principle optical lines from the right camera with respect to left camera's coordinate system. (All calculations from this point will be done in the left camera coordinate system.)

$$l_1 \doteq z_1 = \tan(\theta_r)(x_r - b)$$

$$l_3 \doteq z_1 = \tan(\theta_r - \phi_r)(x_r - b) = p(x_r - b) \quad (5.2)$$

$$l_5 \doteq z_1 = \tan(\theta_r + \phi_r)(x_r - b) = o(x_r - b)$$

The variables m , n , o , p are defined in equation sets 5.1 and 5.2 as

$$\begin{aligned} m &\equiv \tan(\theta_1 - \phi_1) \\ n &\equiv \tan(\theta_1 + \phi_1) \\ o &\equiv \tan(\theta_r + \phi_r) \\ p &\equiv \tan(\theta_r - \phi_r). \end{aligned} \quad (5.3)$$

Each of the critical points **A**, **B**, **C**, **D**, is a vertex on the four sided polygon **ABCD**. Given the location of these points as a function of the stereo variables set **S**, an expression of the area of the stereo sensitive area could be derived.

5.1.1 Solving for the Stereoscopic Region Vertex Points

The first step in finding the area of polygon **ABCD** is to find the physical location of each of the individual points. From figure 5.1, it can be seen that the critical points are defined by the intersection of two of the principle optical rays, one from each of the left and right camera systems. Equation set 5.4 contains the location of each of these critical points in the left camera coordinate system. (See Appendix A for a detailed derivation of each of these points)

$$\begin{aligned}
 \mathbf{A} &\equiv l_2 \cap l_5 = \frac{ob}{o-n}(1, n) \\
 \mathbf{B} &\equiv l_2 \cap l_3 = \frac{pb}{p-n}(1, n) \\
 \mathbf{C} &\equiv l_3 \cap l_4 = \frac{pb}{p-m}(1, m) \\
 \mathbf{D} &\equiv l_5 \cap l_4 = \frac{ob}{o-m}(1, m)
 \end{aligned}
 \tag{5.4}$$

Given these equations, a formula for the area of polygon **ABCD** can be directly obtained from the set **S** of stereo variables.

5.1.2 Derivation of Convergent Stereo Area Equation

Figure 5.2 represents the general configuration of the convergent stereo polygon. Since no special assumptions about the nature of the polygon can be made other than it contains only four sides, the area problem for the polygon must be broken into the sum of the area of two triangles, which can be done quite easily in this case. Equation 5.5 shows the particular break-down that was used in these derivations. (Another combination could have just as easily been used.)

$$\text{Area}_{\mathbf{ABCD}} = \text{Area}_{\mathbf{ABD}} + \text{Area}_{\mathbf{BCD}}
 \tag{5.5}$$

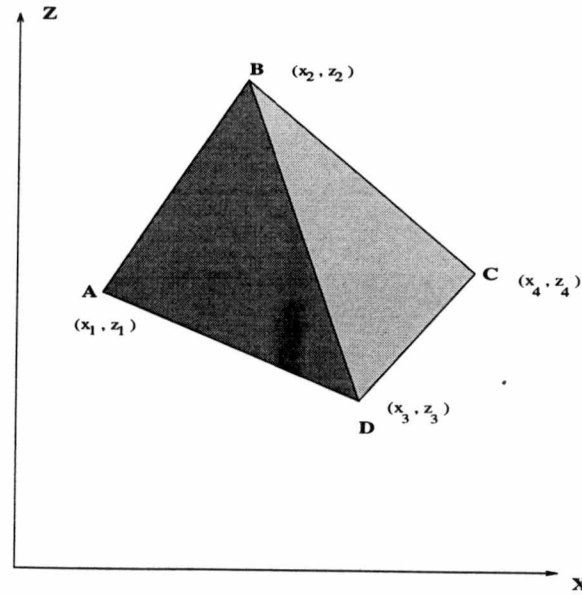


Figure 5.2: **Solution for Area of Quadrilateral:** As the area of the triangle increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.

As the equation shows, the sum of the area of triangles **ABD** and **BCD** equals the area of the polygon **ABCD**. The area of a triangle in a two-dimensional plane is a simple problem, and it can easily be represented in the structure of a determinant as in Equations 5.6 and 5.7 for the triangles **ABD** and **BCD**, respectively.

$$\text{Area}_{\mathbf{ABD}} = \pm \frac{1}{2} \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} \quad (5.6)$$

$$\text{Area}_{\mathbf{BCD}} = \pm \frac{1}{2} \begin{vmatrix} x_2 & y_2 & 1 \\ x_4 & y_4 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} \quad (5.7)$$

Solution of equations 5.6 and 5.7 is a straightforward procedure that results in each of the individual area equations 5.8 and 5.9.

$$\text{Area}_{\mathbf{ABD}} = \left| \frac{onb^2(n-m)(p-o)}{2(o-m)(o-n)(p-n)} \right| \quad (5.8)$$

$$\text{Area}_{\mathbf{BCD}} = \left| \frac{mpb^2(m-n)(o-p)}{2(p-n)(p-m)(o-m)} \right| \quad (5.9)$$

Summation of these individual area equations, as dictated in equation 5.2 results in equation 5.10.

$$\mathbf{A} = \frac{b^2(n-m)^2(p-o)^2}{2|(o-m)(p-n)|} \left[\left| \frac{on}{(o-n)} \right| + \left| \frac{mp}{(p-m)} \right| \right] \quad (5.10)$$

This equation describes the area of the convergent stereo triangle in terms of the stereo configuration variables in the set $\mathbf{S}$. This is an important form of equation 5.2 due to the relationship between area and accuracy in stereo vision. In the next sections equation 5.10 will be used to solve for an *optimal* solution for the set of stereo variables $\mathbf{S}$.

5.2 Isolating Task-Domain Constraints

In [TT95a], a set of task-domain constraints that are formulated in terms of work space geometry is used in conjunction with an area-based criterion function to offer optimal configurations of parallel-axis stereo. Under those simplified conditions, a closed-form solution was easily obtained through analytic means. However, this problem is more complex for two reasons: (1) the area-based criterion function 5.10 does not have an easily obtained derivative. (2) the equation set $\mathbf{S}$ that represents the geometry of a convergent stereo system has five variables. (The parallel-axis problem contained only

two.) In the next section a system of constraints is formulated that will be used to provide logical as well as purposive guidelines for the solution to this problem.

5.2.1 A Circular Work Space Constraint

Figure 5.3 displays a convergent stereo geometry with a circular region inscribed within the area of stereo sensitivity. Each of the configuration variables from set S , as well as the constraint variables are also shown. The focus of this section of the paper is to outline a system of physical and task-domain constraints that will be used to formulate an optimal solution for the variable set S .

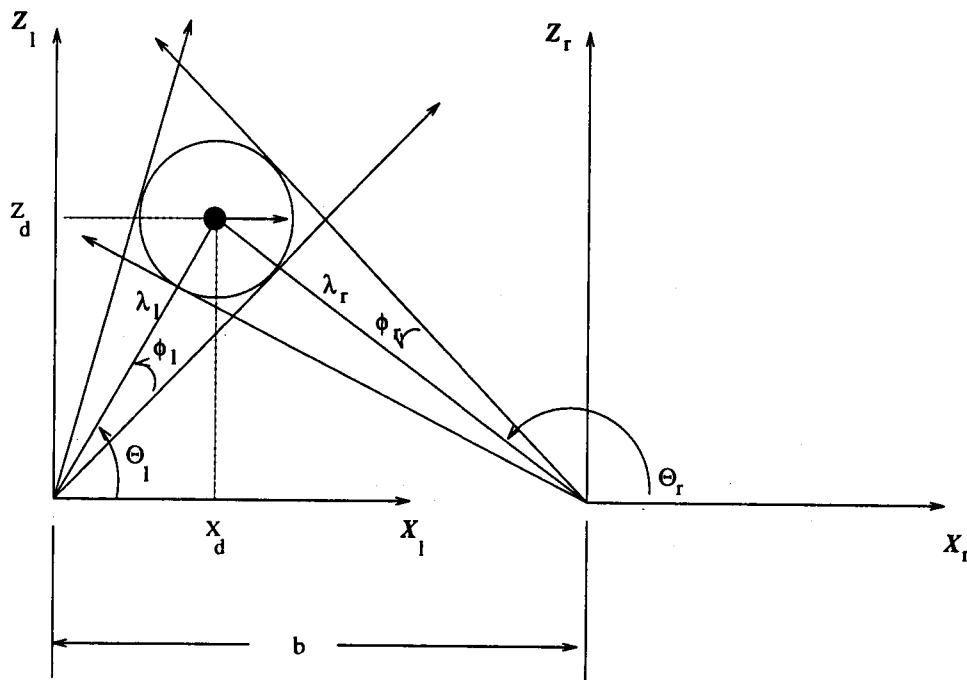


Figure 5.3: **Convergent Stereo Geometry:** As the area of the polygon increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.

The first set of constraints are documented in equation 5.11. These constraints are

primarily derived from some of the basic physical assumptions about convergent stereo geometry. For instance, the convergence angles θ_l and θ_r are permitted to vary between 0 and π radians. Thus, only positive range values can be recognized by the system. Also, the focal length variables ϕ_l and ϕ_r have been constrained to lie on an interval of commonly available focal lengths. This will help to control the convergence of the optimizing function that will be used in the later sections.

$$\begin{aligned}\theta_l, \theta_r &\in [0, \pi] \\ \phi_l, \phi_r &\in [\phi_{\min}, \phi_{\max}]\end{aligned}\tag{5.11}$$

While these constraints are motivated directly from the physical structure of the problem, they are an inadequate basis of constraint to formulate the optimization of equation 5.10. Additional boundaries are required. The addition of a circular region of interest that must be completely contained within the confines of the stereo sensitive area offers the natural form of condition. Motivation for the use of a circular or spherical work area in the arena of robotics-based stereo vision can be found in [TT95a].

Given the general conditions in equation 5.11 and the requirement that a minimum circular area of stereo sensitivity be available to the stereo system, constraints involving the principle optical axes of the convergent camera systems are evident. To begin with the extremal optical axes that bear a direct relationship with the focal length of the system cannot pass through the wedge formed by the set of lines tangent to the circle passing through each cameras optical vertex. This assumption, together with those of equation set 5.11 result in

$$\theta_l - \phi_l - \arctan\left(\frac{Z_d}{X_d}\right) + \arctan\left(\frac{r}{\lambda_l}\right) \leq 0$$

$$\theta_l + \phi_l - \arctan\left(\frac{Z_d}{X_d}\right) - \arctan\left(\frac{r}{\lambda_l}\right) \geq 0$$

(5.12)

$$\theta_r - \phi_r - \arctan\left(\frac{Z_d}{X_d - b}\right) + \arctan\left(\frac{r}{\lambda_r}\right) \leq 0$$

$$\theta_r + \phi_r - \arctan\left(\frac{Z_d}{X_d - b}\right) - \arctan\left(\frac{r}{\lambda_r}\right) \geq 0$$

where

$$\lambda_l = \sqrt{Z_d^2 + X_d^2}$$

(5.13)

$$\lambda_r = \sqrt{Z_d^2 + (X_d - b)^2}$$

and r represents the radius of the circular. The equations in 5.12 effectively prevent the sides of the polygon **ABCD** from intersecting the circle in more than a single point. Naturally, the geometry of the boundary line can be obtained directly from information regarding the radius of the circle and its position relative to either of the two major coordinate frames. Now that an area-based criterion function and a systematic set of constraints have been derived, the principle task of the work becomes *given the geometrical specification of the work space boundaries, what is the "best" configuration for the convergent stereo system to occupy?*

5.3 Nonlinear Constrained Optimization for Convergent Stereo Configurations

Equation 5.10 is a nonlinear equation in five unknowns. The set of task-domain constraints in equations 5.11 and 5.12 are not sufficient conditions for the existence of a unique solution for the variable set S . Therefore, a minimum in the criterion function represented by equation 5.10 will be sought under the conditions and constraints documented in equations 5.11 and 5.12. At this point the construction of the problem, there is an implied assumption that the area of polygon $ABCD$ has a direct relationship with the accuracy of the range data produced by a convergent stereo algorithm operating on it. However, this assumption will be justified in the next section.

A Sequential Quadratic Programming (SQP) algorithm that is directly available in the MATLAB optimization toolbox [Gra95] was used as the optimization algorithm that would seek a constrained minimum in equation 5.10. From figure 5.3, it seems that the minimum area should occur when the intersection of each of the extremal principle optical rays is a single tangent point to the circular work space. Also, this formulation would require the optical axis of each camera to intersect at the center of that circle. Convergence of equation 5.10 should settle near those points which have an easily obtained analytic solution.

In order to provide some broader applications of this theory two specific types of convergent stereo systems will be examined. First, equation 5.10 will use the baseline variable, b as a constant. This will have the effect of reducing the unknowns to four variables and accordingly a simplification of the problem. This case is motivated by the existence of a large number of convergent stereo machines that do not support a moving

baseline [FC93,MT94a,Tso94]. Lastly, the case of convergent stereo with adjustable baseline will be studied. This case is the most general form of equation 5.10 and will result in the most accurate range data possible. As a consequence of the advantages dynamic baseline configurations offer, many new stereo machines have this attractive feature [PE93,Chr93,KFS88].

5.3.1 Static Baseline Configurations

As mentioned, static baseline configurations change the structure of equation 5.10 to the minimization of an equation in four unknowns instead of five. Also, it was mentioned that there is considerable support provided in the literature for such an endeavor. The problem is structured as follows:

- Input task-domain constraints.
- Find S , with baseline held constant, that minimizes equation 5.10.

Specification of task-domain constraints essentially involves choosing three things: (1) the radius of the circle, r (2) the geometric position of the center of the circle with respect to the fixed coordinate frame, the variables Z_d and X_d . From these the system of constraints in equation 5.12 can be computed. (3) the *a priori* selection of baseline distance based on a criterion of the designers choice. Appendix A contains the MATLAB function file that is necessary for input to the algorithm and a sample specification of these three requirements.

Interestingly, the solution of the posed problem by MATLAB turns out to be the same solution that was posed earlier by observation of the geometrical construction of figure 5.3. If, for the moment, it is assumed that this is always the optimal solution or

more appropriately, at least at this point in the development, near the optimal solution, the seed point and possibly the solution can be obtained from purely analytic means.

The solution set is

$$\begin{aligned}\phi_l &= \text{asin}\left(\frac{|X_d|}{\lambda_l}\right) \\ \phi_r &= \text{asin}\left(\frac{|X_d - b|}{\lambda_r}\right)\end{aligned}\tag{5.14}$$

$$\theta_l = \text{atan2}(Z_d, b)$$

$$\theta_r = \text{atan2}(Z_d, X_d - b).$$

If this solution analytic solution violates the hard limits on focal length and baseline that are used to constraints the optimization, the MATLAB routine should be used to provide the nearest design point that lies within the feasible region.

5.3.2 Dynamic Baseline Configurations

So far the formulation of the problem has been relatively straight forward. The basic assumptions of static baseline have affected a tremendous reduction in the complexity of the problems posed thus far. For example, in static baseline configurations the coordinate systems remain fixed with respect to the task-domain specifications, e. g. X_d and Z_d remain constant with respect to left and right camera systems. In a situation where baseline is allowed to move, there is an implied movement of one or both of the camera systems. Thus, a reformulation of the problem is required. Two specific cases are posed: (1) only one of the camera systems is allowed to move; the other remains fixed with respect to some absolute reference. (2) both camera systems are allowed to

move with respect to an absolute reference. They are referred to as *Constrained* and *Unconstrained* baseline movements, respectively. They require the solution of a more complex problem due to the dynamics involved in the calculations of the task-domain constraints with respect to the camera coordinate frames.

Constrained Baseline Movement

Even though constrained baseline movement is not the most general case of baseline translation, it was chosen for study because it maintains a sympathetic relationship with equation 5.10 in one of its forms. In this model, the baseline variable, b , is changed by moving the right camera with respect to a fixed coordinate system at the left camera system. (Notice: this will have a different affect on the system geometry than moving the other set of axes.) Also, some active vision platforms may have similar kinematic construction where a single linear actuator is used to provide baseline adjustment. Figure 5.3 displays the two-dimensional kinematic composition of such an arrangement.

Unconstrained Baseline Movement

In order for the convergent stereo system to extract the maximum "quality" of range information from a scene, the geometry of the baseline must be allowed to vary in an unconstrained fashion. This implies that the left and right camera systems must be allowed to move in an independent fashion so long as the condition that $b_l \leq b_r$ is preserved. See Figure 5.4 for the geometry of this model.

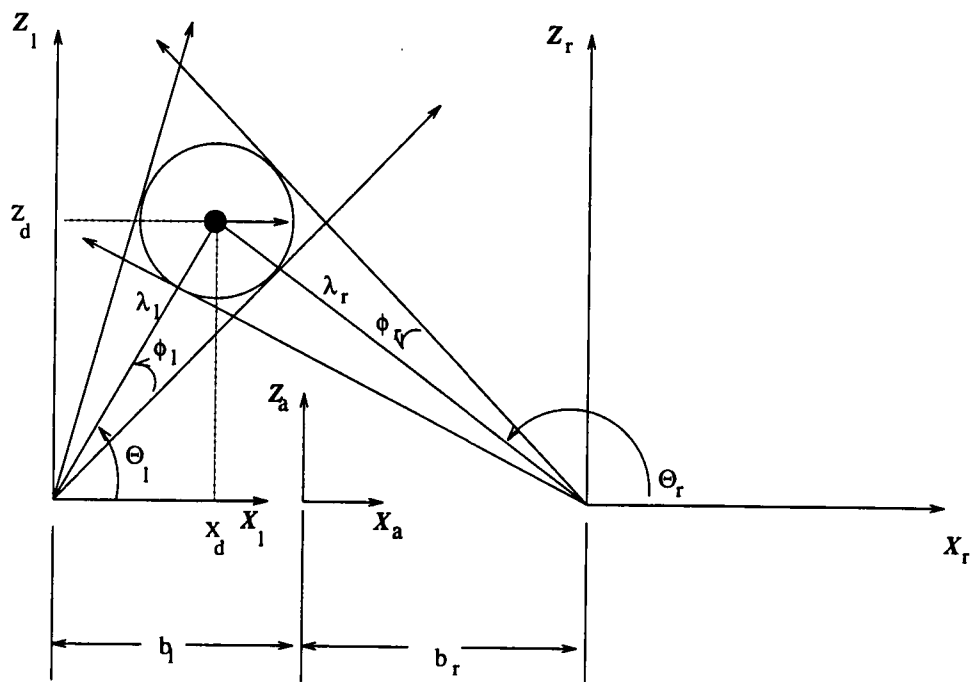


Figure 5.4: **Convergent Stereo Geometry with Unconstrained Baseline Translation:** As the area of the polygon increases more objects are included in the region of binocularity. However, the accuracy of the range estimates in that triangle decreases.

5.4 The Convergent-Axis Stereo Synthesis (CASS) Algorithm

The Convergent-Axis Stereo Synthesis (CASS) algorithm can acquire a closed-form solution if the allowable range of f is within the solution space. If this is true, a closed-form solution for the static baseline problem can be obtained from equation set 5.14. The convergence of both of the dynamic baseline systems is a function of a more complex interaction of baseline and focal-length. This interdependence will prevent a closed-form solution from being easily realized. In these cases a seed baseline should be selected and then the static baseline subproblem should be solved to offer a good approximation to an initial solution. Figure 5.4 outlines the basic structure of each of the three convergent stereo algorithms.

5.4.1 Specification

At this point, the choice of geometries of the workspace is limited to the specification of the parameters of a circle and its geometrical distance from a fixed camera coordinate frame. It is a good idea to outline, before selecting the workspace geometry, the general goals of the system. If the system is to be used in its principal role as an *identification* tool, a large workspace geometry which emphasizes maximizing the stereo sensitive area will be desirable. However, if the system is to be used in a *localization* task where the accuracy of the recovered data is paramount, the workspace geometry should be made as small as possible. This competing relationship between workspace geometry and range accuracy is called the identification/localization tradeoff.

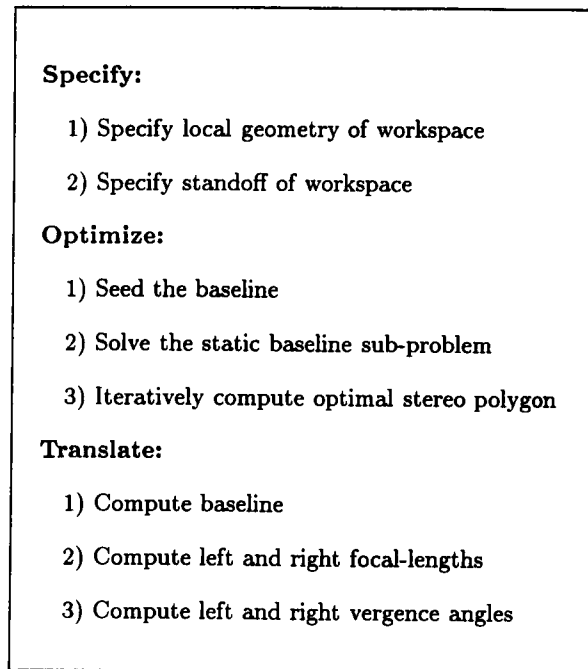


Figure 5.5: **Convergent-Axis Stereo Synthesis (CASS) Algorithm:** Systemic constraints are mapped into unique stereo geometries using an iterative minimum area difference technique.

5.4.2 Optimization

This is the most critical stage in the algorithm. And must be performed differently in each of the three algorithms. It can be trivial in the static baseline convergent stereo applications, and can also be computationally prohibitive in the more complex dynamic baseline cases. Regardless of which algorithm is used, however, the concept is essentially similar. The appropriate subset of S is located after specification of the work space constraints. Some of the variables in S do not represent directly camera parameters of the design; rather, they are an intermediate representation that is better suited for geometrical study.

Static Baseline Optimization

Static baseline optimization can be handled in a very simple manner. Simply solve equation 5.14 and see if those solutions lie within the allowed boundaries 5.11 and 5.12. If they do then that solution is optimal and can be used directly. If computed focal-length is outside of the feasible reason, simply chose the closest permissible focal length that your camera system will permit. If the vergence angles have been computed out of boundaries, there is probably an error in the task-domain constraint specification or the specification is geometrically not suitable for convergent applications, i. e. vergence angles that are 180 degrees out of phase. The formal steps are:

- Solve equation 5.14 after inputting task-domain constraints.
- Use the those numbers as a seed point to the algorithm.

Constrained Baseline Optimization

Constrained baseline optimization occurs in a fashion similar to the static baseline case. In fact, the only difference between them is that the baseline is allowed to vary over the range of physically realizable areas: $b \in [b_{\min}, b_{\max}]$. This problem can be studied as a generalization of the previous static baseline formulation. If an initial point for the baseline is chosen from the state of the previous configuration or an initial guess from the user, then the problem is reduced to the static baseline sub problem. Since the solution to this problem is known, solving it will provide the constrained baseline algorithm with an good initial guess for the remainder of variables from set S . The formal steps are:

- Select a seed value for the baseline, b .

- Solve the static baseline sub problem to find the other seed variables.
- Optimize the full version of equation 5.10.

Unconstrained Baseline Optimization

The optimization of the unconstrained baseline system will offer the most stable geometrical configuration that a convergent stereo system could offer. It will involve the minimization of an area criterion function in six variables whose formulation will be significantly different from equation 5.10. However the overall procedure will consist of:

- Select a seed value for the baseline, b_l and b_r .
- Solve the static baseline sub problem to find the other seed variables.
- Optimize the full version of unconstrained baseline area function.

This configuration will allow the convergent system to bring its version angle (see figure 5.6) to values closer to a symmetric vergence system. Such configurations offer the premium accuracy values of any of the three systems thus described.

5.4.3 Translation

Convergent-axis stereo is completely specified by a baseline, left focal length, right focal length, the vergence angles. The final stage of the CASS algorithm is a translation of the geometrical design parameters S into an actual set of variables that can be feed to a controller unit or computer simulation environment. Clearly, the baseline is either explicitly set prior to the optimization, as in the static baseline problem, or it can be directly obtained from S . The same is true for the vergence angles θ_l and θ_r . Therefore,

the only translation involved in these algorithms is the conversion of the left and right field-of-view variables into their equivalent focal lengths. This is done with $f = \frac{\rho}{\tan(\phi_c)}$ where c represents the left or right camera and ρ is a variable measuring the dimension of the image plane.

5.5 A Geometric Description of the Optimality

The key factor in each of the previous arguments was the assumption that the area of polygon **ABCD**, in its minimum form with respect to equations 5.11 and 5.12, has a direct relationship with the accuracy of the derived range values. In [TT95a] a geometrical argument borrowed from [ABG89] was used to establish this relationship. However, Christensen relates in [Chr93] that the relative accuracy for measurements taken along the *cyclopean* axis of both symmetric and asymmetric convergent stereo systems is a function independent of baseline and focal length is not included in his model. He goes on to say that only a high degree of angular resolution is required to satisfy the accuracy requirements of his model shown in 5.6. He motivates the need for high accuracy in angular resolution due to the fact that his vergence angles are close to ninety degrees (an ill-conditioned region in his formulation). However, he is wrong on two accounts: (1) if the baseline had been adjusted to account for the depth of the view point, the vergence angles would not be close to singular and the problem would be better conditioned (2) no one really uses vergence stereo in "point" mode; it is too inefficient. Using vergence in a region mode involves the addition of real disparity in a region of depth neighboring the intersection of the optical axis. These disparities, and consequently, the accuracy of the measurements are maximized in the

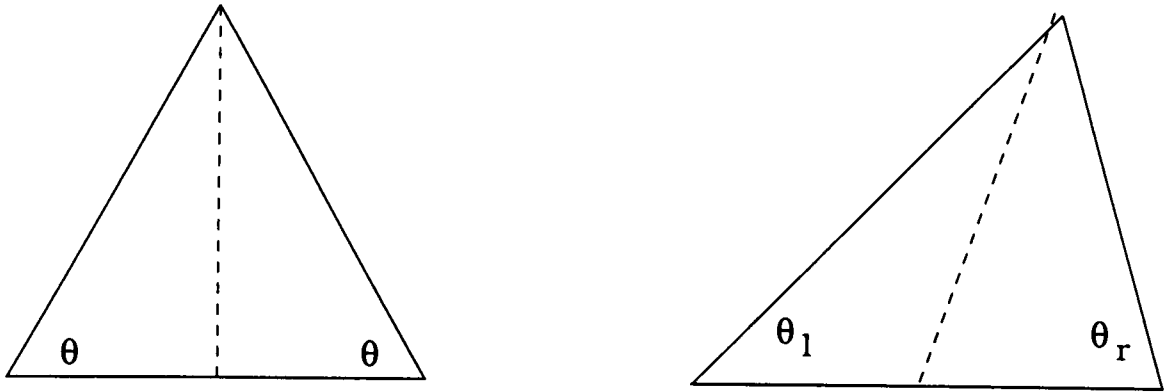


Figure 5.6: **Convergent Stereo Model:** A typical convergent stereo model relies on pure angular resolution to provide accurate depth measurements when the real problem is the integrity of the geometry.

same manner as parallel-axis stereo, the baseline and focal length product. Given this, some general assessments about the relationship between the task-domain constraints and the accuracy of the resultant depth measurements can be made.

- The focal lengths of each camera system, ϕ_l and ϕ_r , will be a function of the radius of the circle, r and its distance from the camera frame.
- The baseline variable will be a function of the depth and the type of baseline motion allowed by the model.
- The vergence angles, θ_l and θ_r , will be simple functions of the relative distance between each of their coordinate frames and the center of the circle in those frames.

5.6 Conclusions

A framework for the generation of convergent-axis stereo geometries was presented. This framework was applied to two separate, but highly related problems in convergent

stereo: (1) *static baseline convergent stereo* (2) *dynamic baseline convergent stereo*. The introduction of a system of physical and task-domain constraints was used in a nonlinear optimization function to arrive at a suitable set of stereo configuration variables. The solution employed a cost function based on the inherent accuracy/area relationship in a convergent stereo system. As a consequence of the structure of the problem, a closed-form solution can be used instead of the iterative optimization. The existence of this relationship allowed the construction of the Convergent-Axis Stereo Synthesis or CASS algorithm. Using CASS, a designer can translate work space and standoff requirements into an optimal design or configuration for a convergent stereo system.

CHAPTER 6

Calibration for Integrated Active Stereo

6.1 Introduction

Accurate calibration of camera systems is an essential first step in a computer vision system requiring three-dimensional measurements [WCH92, IPS87, HN91, Nom92]. There are two general philosophies for attacking this problem. The first methodology concentrates on experimentally determining internal camera parameters [WCH92]. In this scheme, knowledge of first-order camera parameters and possibly distortion models work through the perspective transformation to uncover a relationship between image space and three-dimensional space. The second methodology, however, requires specific knowledge of external parameters (i.e. a calibration scene geometry) and uses this to provide a derived transformation between image and world space. Here, the principle technique uses a linear least-squares algorithm to extract this transformation empirically from the environment [IPS87, HN91]. Calibration is achieved by fixing the scene geometry to a known location, and deriving these transformations from it. Inherent in the recovery of the transformation is a matrix inversion process. In many cases, this inversion may produce a stable result; however, inaccuracies in modeling and measurement and random perturbations can introduce uncertainty that could result in loss of rank in the matrix. The attractiveness of this method stems from its ability to produce accurate calibration without knowledge of the elusive internal camera parameters.

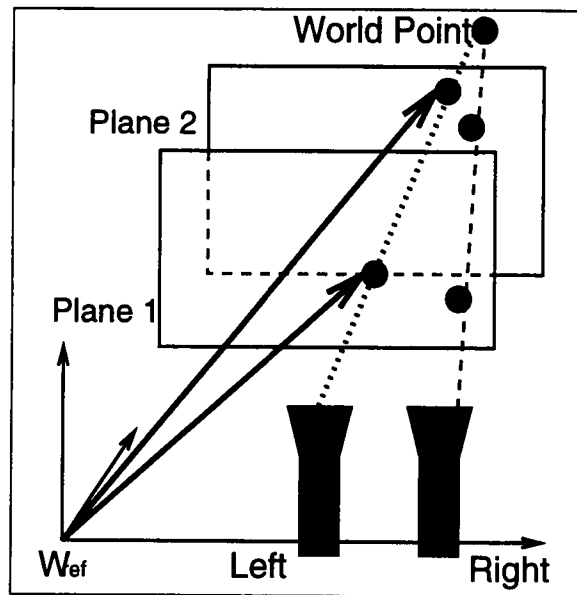


Figure 6.1: **Stereo Two-Planes Calibration:** Depicting imaging vectors from each stereo camera pair, and their corresponding intersections with the calibration planes.

The Two-Planes method as described by Izaguirre [IPS87, Tsa89] was chosen for these reasons.

Conceptually, Two-Planes calibration originates from the idea that a line from a reference point, W_{ref} , to a point in a plane will intersect that plane at a unique point. When an additional plane is added to the scene, as in Figure 6.1, a vector emanating from the imaging plane in W_{ref} space is described. The power of this concept becomes clear. If the transformation between points located on two known planes in the calibration scene and a world reference frame is known, a point in three dimensional space can be described by a vector in W_{ref} . Solving for this transformation involves many interesting numerical phenomena. Of particular importance is the matrix inversion inherent in the either of the proposed solutions. Study of the accuracy and stability of this process via the singular value decomposition will outline some guidelines for implementing two-

planes in a “real” scenario.

With respect to *Two-Planes* calibration, several separate analyses have been performed that relate residual error and metric uncertainty into the accuracy of recovered calibration targets [DK93, IPS87, Tha93]. These studies, however, neglect the propagation of uncertainty in these intermediate points throughout the vector operations required to reconstruct a three-dimensional point from them. Therefore, given a set of calibration parameters and an associated set of assumptions made with respect to randomness in the target geometry, we wish to generate a valid number that describes on the average the maximum uncertainty associated with a given reconstruction. Figure 6.1 presents a typical calibration scenario showing both intermediate and reconstructed three-dimensional points.

6.2 Calibration Mathematics

It was stated that the exact location of two planes in the world coordinate frame must be known. Therefore for each world plane, a transformation that describes the geometrical relationship between points in the image plane and points in the each calibration plane will be derived. Using standard linear interpolation a general form of this transformative process is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_i = \mathbf{A}_{nci}(x^n, x^{n-1} \times y, \dots, y^n, \dots, x, y, 1)^T, \quad (6.1)$$

where i identifies the particular world pane; n is the order of the polynomial fit; $(XYZ)_i^T$ are the coordinates of a world point from plane i ; $(x, y)_c$ are the coordinates of the same

world point in the image plane of camera c ; and A_{nci} is the interpolation matrix for the plane i and the camera c . Using matrix notation where there are N points of interest, this expression may be written as

$$\mathbf{P} = \mathbf{L}\mathbf{A}_{nci}^T. \quad (6.2)$$

Here, $\mathbf{P}$ is a $N \times 3$ matrix of world points, and $\mathbf{L}$ is a $N \times (n+1)(n+2)/2$ polynomial interpolation matrix. In calibration $\mathbf{P}$ and $\mathbf{L}$ are known, while $\mathbf{A}_{nci}$ is usually solved employing one of two techniques outlined in the next section.

6.2.1 Derivation of $\mathbf{A}_{nci}$

Given a matrix of world coordinate points from plane i with respect to W_{ref} and their corresponding projections into camera c , Equation 6.2 is left with one unknown. Theoretically, the linear solution should exist. In most cases, the matrix $\mathbf{L}$ is not square; therefore, a simple premultiplication by its inverse is prohibited. Two useful methods for solving problems of this type are the *Moore-Penrose Pseudo Inverse* (MPPI) and *Singular Value Decomposition* (SVD). The least squares solution via the pseudo-inverse as described by Izaguirre [IPS87] produces

$$\mathbf{A}_{nci}^T = (\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{P}, \quad (6.3)$$

while the SVD for *overdetermined* systems generates

$$\mathbf{A}_{nci}^T = \mathbf{V}[\text{diag}(1/\sigma_i)](\mathbf{U}^H\mathbf{P}), \quad (6.4)$$

where

$$\mathbf{L} = \mathbf{U}[\text{diag}(\sigma_i)]\mathbf{V}^T. \quad (6.5)$$

Selection of the solution mechanism will have significant physical and numerical consequences.

6.2.2 Selection of Solution Algorithms

The first and most important decision in the solution of any linear least-squares estimation problem is the choice of one of these two methods. To begin, equation 6.4 provides a unique solution only for the overdetermined system, i. e. $N \geq (n + 1)(n + 2)/2$, while equation 6.3 has a unique solution, independent of n , for $N \geq 6$ [Pre88, IPS87]. Thus solution of the *normal* equation represented in equation 6.3 will theoretically allow arbitrary polynomial interpolation order n as long as at least six calibration points are present. Solution via the SVD, however, introduces a limitation on the number of terms any given interpolation polynomial can have. More precisely, a two dimensional (a plane) polynomial of order n requires $(n + 1)(n + 2)/2$ coefficient terms and consequently the same number of calibration points [TG74]. This constraint will become more severe as higher order interpolations are desired. Aside from these basic physical implications, selection of either of the equations necessarily introduce some numerical constraints. Naturally both methods require that L be a nonsingular. Assuming this, Strang relates that it is always better to avoid solving a normal equation [Str86]. This law results from the property that the eigenvalues of $(L^T L)$ are the square of the eigenvalues of L [Str80, ND88]. Therefore, given a sufficient number of calibration points the singular value decomposition should be used before the Moore-Penrose pseudo inverse. Table 6.1 illustrates the interrelationship of the eigenvalues as expressed through the condition numbers of the transformation matrices (P contained 80 points at a standoff distance of 124.46 centimeters).

Table 6.1: **Typical Condition Numbers vs. Interpolation Polynomial Order:**

Column 1 reports the order of the interpolation polynomial. Column 2 contains condition numbers for the singular value decomposition while column 3 pertains to the Moore-Penrose pseudo inverse.

Order	SVD	MPPI
1	1280	1.6×10^6
2	1.4×10^6	1.9×10^{12}
3	1.6×10^9	2.6×10^{18}
4	1.9×10^{12}	4.0×10^{24}
5	3.1×10^{15}	9.8×10^{30}

6.2.3 Selection of Basis Polynomials

Once an algorithm for the solution of a linear least-squares estimation problem has been chosen, the extremely difficult task of choosing the correct basis polynomials for equation 6.1 must be tackled. The accuracy of the calibration hinges on the ability of the system to recover the precise location of individual targets in a calibration plane. A method for testing the accuracy of an interpolation process is to measure the relative difference between the physical location of the known data and the values as derived by the interpolation process. This difference is called the *residual* error in the transformation. Measuring it will indicate how well a particular transformation approximates the known data. As the order of the basis polynomials is allowed to increase the residual error in the point recovery decreases as a nonlinear function of n . Figure 6.2 displays this relationship for three different calibration planes. Observation of this information alone would seem to indicate that the largest interpolation order is the most desirable;

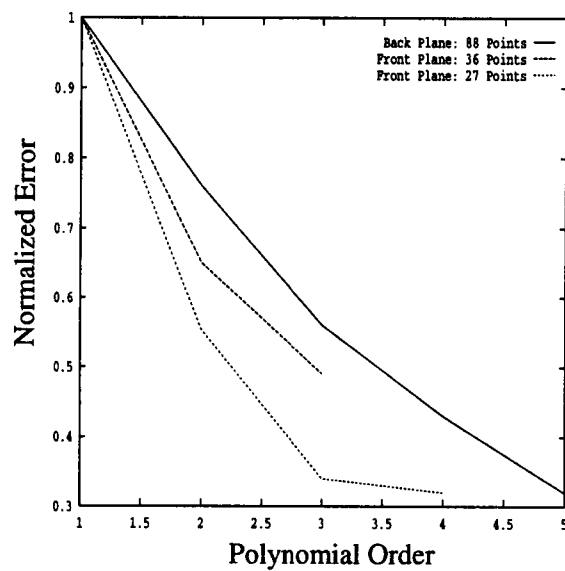


Figure 6.2: **Normalized Error Residue vs. Interpolation Order:** Three different physical calibration planes were implemented with standoff distances of 124.46, 83.82, and 45.72 cm.

however, the stability of the transformation process must also be examined.

Another important issue in basis polynomial selection is *singularity*. Singular value decomposition theory states that every $p \times q$ matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \quad (6.6)$$

where $\mathbf{U}$ is $p \times p$ and unitary, $\mathbf{V}$ is $q \times q$ and unitary, and $\mathbf{\Sigma}$ is $p \times q$ and diagonal in that $\langle \Sigma_{ij} \rangle = 0$ unless $i = j$, in which case we write $\langle \Sigma_{ii} = \sigma_i \rangle$, real and nonnegative [ND88]. The values, σ_i , are called the *singular values* of $\mathbf{A}$, and the *rank* of $\mathbf{A}$ is the number of nonzero singular values of $\mathbf{A}$. It must be shown that the inverse of the matrix $\mathbf{A}$ is stable (i.e. $\mathbf{A}$ does not lose rank). If the matrix $\mathbf{A}$ is square, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$ are likewise square matrices of the same size. Through the decomposition of $\mathbf{A}$, any instability in $\mathbf{A}^{-1}$ may be exposed. The matrices $\mathbf{U}$ and $\mathbf{V}$ are square and unitary; therefore, the inverses are their hermitian transposes. The matrix $\mathbf{\Sigma}$ is diagonal and its inverse may be obtained by taking the reciprocal of the individual σ_i . Press [Pre88] formalizes this as

$$\mathbf{A}^{-1} = \mathbf{V}[\text{diag}(1/\sigma_i)]\mathbf{U}^H, \quad (6.7)$$

and states that the instability will occur only if one or more of the σ 's becomes zero, or $\mathbf{A}$ loses rank. Because this theory supposes that the numerical tools have infinite resolution, modifications must be made to support the finite precision of a digital computer. Therefore, any inverse will not be stable if the reciprocal of one of the σ_i 's is so small that its value is dominated by roundoff error and therefore unknowable [Pre88]. This is expressed formally as the ratio of the largest eigenvalue, or σ_i , to the smallest, and is called the *condition number*. If the condition number is infinite, then $\mathbf{A}$ is singular. If the number is large with respect to the floating point precision of the machine, the

A is *ill-conditioned*. The IEEE 754 standard for floating point arithmetic outlines the smallest increment that is representable by a given floating point number as a *long-double* without loss of precision is 1×10^{-19} [Pla92]. Comparison of this with value with condition number of each transformation will indicate the integrity of the derived transformation. Table 6.1 indicates that matrix condition is also a nonlinear function of polynomial order.

Two important constraints for basis polynomial selection have been outlined. First, the transformation must produced a set of data that are known to be accurate within a certain range. This range will form the basis of all future analysis with respect to error in the stereo reconstruction. Simultaneously, the transformation matrices must not be singular with respect to the floating point precision of the computing machinery. An algorithm for basis selection is governed by two competing functions of n . Large n gives more accurate results at the expense of stability, and visa-versa.

Strategies for Improved Stability

Matrix condition is directly related to the format of the data stored in L . Each row vector from L is a polynomial constructed from the basis template using a set of image coordinates as its arguments. Traditionally, these image plane coordinates are represented as integers. In the case where a 512×512 pixel array is employed, representation of image coordinates by integers will introduce coefficients in L that span many orders of magnitude. Stark warns against this type of representation while Depiero claims that by normalizing the image coordinates the span of coefficients in $L^T L$ was reduced from 21 to 8 orders of magnitude in some cases [Sta70, DK93, TT95b].

Ideally a first order basis polynomial should synthesize an accurate estimate of the

imaging process following from its relationship with the perspective transformation. Lens distortion, however, can introduce highly nonlinear affects that make a first order model inappropriate near the boundaries of the imaging plane; consequently, higher order modeling is necessary in applications where camera motion is prohibited or the entire image structure is to be exploited. Goodman relates that most lens aberrations can be modeled as second or third order affects with good results [Goo68]. This supposition is further fortified by consideration of the calibration data that is stored in $\mathbf{P}$ which lies in a single physical plane, i. e., it is two-dimensional information. A good rule of thumb is never to use a basis polynomial with order greater than three unless it is absolutely necessary.

Synthesis of Calibration Polynomials

In this section, some polynomials of various order are are loaded with test data from one of the six calibration planes. These polynomials are are analyzed using both the SVD and MPPI algorithms. Based on the aforementioned matrix condition checks, the calibration system returns a binary vector indicating the significant polynomial terms. New interpolation polynomials can be generated by cross-referencing this vector with the original polynomial and removing the appropriate terms with minimal loss in accuracy. Tables 6.2, 6.3, and 6.4 display the suspect terms and their corresponding synthesis flags. All of the lower order terms (not displayed) were found to be significant for both the MPPI and SVD solutions.

Table 6.2: **Synthesis of Significant Polynomials(High Order)**: Constraints on matrix condition and residual error produce a vector that describes whether a particular polynomial coefficient can be zeroed. The first column indicates from which camera the data was obtained, while the second column indicates the algorithm.

Camera	Alg	x^5	x^4y	x^3y^2	x^2y^3	xy^4
1	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1
2	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1
3	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1

6.3 Monocular Error Analysis

Any analysis of matrix condition must include a study of the effects of perturbation on it. If small changes in the measured data induce unreasonably large changes in the derived transformation, it may prove to volatile to use in a practical setting. Table 6.5 displays the condition numbers from three of the six data sets. The measured data was disturbed in both the image and calibration space. Image space points were violated with uniformly distributed noise in the interval $[-1, 1]$ (pixels), while calibration targets were assumed to have an uncertainty of ± 0.015 inches in all directions. Simulations were run for 1,000 and 1,000,000 iterations. At each step the conditions numbers were compared with a reference condition number to check for unusual problems (non were reported!), and an average condition number was synthesized. Unfortunately, it can not

Table 6.3: Synthesis of Significant Polynomials (Mid Order).

Camera	Alg	y^5	x^4	x^3y	x^2y^2	xy^3
1	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1
2	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1
3	MPPI	0	0	0	0	0
	SVD	1	1	1	1	1

be said from this type of analysis that the transformation matrix is nonsingular. For any given data set, the reciprocal of its condition number measures (with respect to a given vector norm) how far it is from the nearest singular matrix. Given that the magnitude of the induced perturbations is considerably larger than the best conditioned matrix, no definite guarantees can be given that will assure convergence. The only solution becomes to remove the polynomial terms that give are responsible for realizing large condition numbers.

6.3.1 Geometric Interpretation of Monocular Error

It is desirable to extract some type of measure from the residual error matrices to describe in a meaningful geometric sense the errors inherent in the least-squares approximation. We turn to models from linear algebra of the various norms of a matrix and try to interpret some meaning from them. The most common norms on a $p \times q$

Table 6.4: Synthesis of Significant Polynomials(Low Order).

Camera	Alg	y^4	x^3	x^2y	xy^2	y^3
1	MPPI	0	1	1	1	1
	SVD	1	1	1	1	1
2	MPPI	0	1	1	1	1
	SVD	1	1	1	1	1
3	MPPI	1	1	1	1	1
	SVD	1	1	1	1	1

matrix A are the *One*, *Two*, and *Infinity* norms which are defined as

$$\|A\|_1 = \max_j \sum_{i=1}^p |a_{ij}|, \quad (6.8)$$

$$\|A\|_2 = \max \sigma_{ii}, \quad (6.9)$$

and

$$\|A\|_\infty = \max_i \sum_{j=1}^1 |a_{ij}|. \quad (6.10)$$

respectively. Once these standard tools are applied to the residue matrices, some useful geometric information is derived, as shown in Table 6.6. The first column in the table represents the maximum absolute value of the elements in the matrix. Geometrically, this represents the limiting uncertainty that a single point can have in one dimension. If we assume that this error exists in each element of a three-dimensional coordinate vector obtained by the transformation, the distance formula will produce the radius of a ball centered at that particular coordinate vector. This measure is the maximum error, per recovered point, that the transformation coupled with quantitative assumptions

Table 6.5: **Effects of Perturbation on Condition:** Simulations where run for 1000 and 1,000,000 iterations respectively.

Camera	Order	10^3	10^6
1	1	1280	1280.7
	2	1.403×10^6	1.403×10^6
	3	1.617×10^6	1.618×10^9
	4	1.977×10^{12}	1.997×10^{12}
2	1	1935	1935
	2	3.365×10^6	3.364×10^6
	3	6.391×10^9	6.389×10^9
	4	1.352×10^{13}	1.351×10^{13}

concerning the calibration geometry will acquire. Column three of Table 6.6, the infinity norm, provides a less severe estimation of this uncertainty sphere by maximizing the sum of the absolute values of the rows. The One norm, from column two, describes a cumulative error effect in one-dimension, such as the sum of errors in the x direction. This is clearly a function of the number of calibration points and not descriptive of error in the recovered geometric structure.

The presence of ground truth information in calibration permits the derivation of accuracy information in lieu of repeatability; consequently, the element norm or the infinity norm (from table 6.6) provides direct measure of the accuracy of two planes calibration. Given several sets of such data, computation of the variance among one of the columns of Table 6.6 will measure the precision of the calibration. Table 6.7 contains the accuracy and precision information using both element and infinity norms.

Table 6.6: **Geometric Error Analysis:** Six sets of test data are analyzed with three different norm functions (mm).

Camera	Element	One	Infinity
1	0.406	5.257	0.670
2	0.177	1.270	0.254
3	0.711	3.733	0.889
4	0.812	26.94	1.143
5	0.711	24.56	1.066
6	0.812	27.71	1.092

Table 6.7: **Geometric Error Analysis:** Derivation of accuracy and precision. (mm)

Norm	Accuracy	Precision
Element	0.594	0.2351
Infinity	0.852	0.3109

It must be pointed out that these measures were composed of data taken at a range of calibration distances (45.72-124.46) cm. Thus, the information can be taken as an estimate for the error that will be realized in calibration distances near this interval. For instance, if Equation 6.1 is applied to a set of image coordinates an estimate of the uncertainty in the three dimensional point can be obtained from an appropriate selection of the values of Table 6.7. It would be useful at this point to determine the sensitivity of the stereo triangulation algorithms to the residual error estimates contained in Tables 6.6 and 6.7. (All stereo analysis is performed using information obtained from a second order calibration polynomial.)

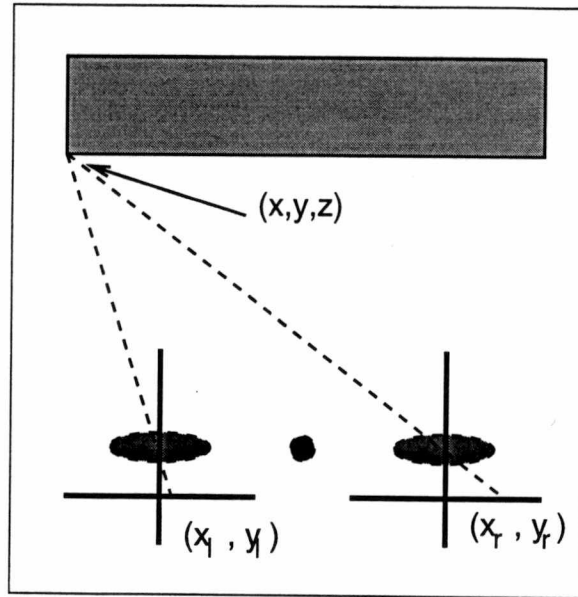


Figure 6.3: **Basic Geometry:** Stereo vision system.

6.4 Binocular Error Analysis

Fu acknowledges that the mapping of a three-dimensional scene onto an imaging plane is a many-to-one transformation [FGL87, GW87]. Binocular stereo is a well known technique for recovering the missing information. These techniques involve the use of differential information gained from analysis of two separate camera views of the same three-dimensional scene to reconstruct the original point. Figure 6.3 and Equation Set 6.11 illustrate the intuitive concepts behind this class of techniques [Hor89].

$$\begin{aligned}
 x &= b \frac{(x_l + x_r)/2}{x_l - x_r} \\
 y &= b \frac{(y_l + y_r)/2}{x_l - x_r}, \text{ and} \\
 z &= b \frac{f}{x_l - x_r}.
 \end{aligned} \tag{6.11}$$

While these first order relationships are relatively simple and do not lend themselves

to an accurate recovery of three-dimensional depth information, they do, however, reveal the nonlinear relationship in the mapping processes. If it is desired to translate uncertainty in the image space projections into a measure of error in their recovered locations, propagation of a valid error signal throughout these simple equations should be sufficient. However, in a more sophisticated scenario these equations are replaced with a more adequate representation; consequently, the presence of nonlinearities in the context of a specific calibration algorithm require a more complex analysis.

Given a stereo set of cameras viewing a single point in a three-dimensional scene, an image processing system capable of producing a set of corresponding image space points, one from each camera, becomes the input to Equation 6.2. After application, a four member set of three-dimensional points that represent the projection of the scene point onto each of the calibration planes is generated (see Figure 6.1). This information defines two vectors, one from each camera, that are oriented toward the scene point, and the intersection of these vectors defines that point. Figure 6.4 depicts a complete vector diagram of a scene point, its projections onto two separate calibration planes for a binocular camera system, and a complete vector description of their relationships. Problems will result due to the presence of nonlinearities, residual and measurement error, and spatial quantization affects that prohibit the intersection of these idealized vectors. Duda and Hart propose locating the scene point one the midpoint of the line of minimum distance between the two vectors $\vec{r}_1$ and $\vec{r}_2$. Two sets of coordinates vectors $(\vec{P}_{11}, \vec{P}_{12})$ and $(\vec{P}_{21}, \vec{P}_{22})$, returned from application of Equation 6.2 for both calibration planes join to form $\vec{r}_1$ and $\vec{r}_2$ through

$$\begin{aligned}\vec{r}_1 &= \vec{P}_{21} - \vec{P}_{11} \\ \vec{r}_2 &= \vec{P}_{22} - \vec{P}_{21}\end{aligned}\tag{6.12}$$

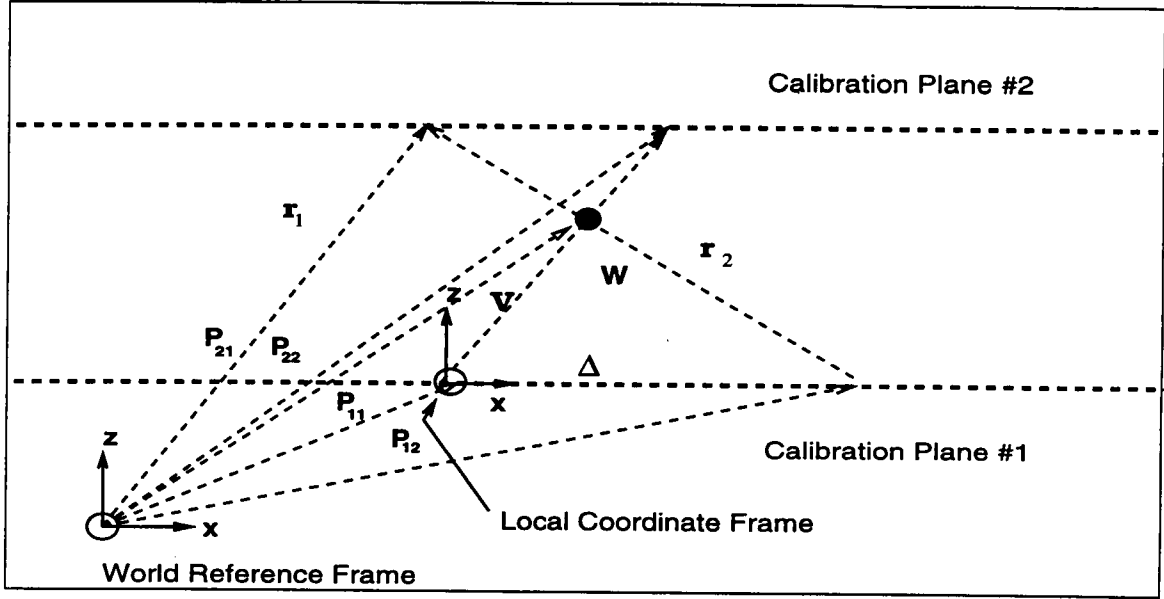


Figure 6.4: **Complete Geometrical Description:** Solution to point of minimum distance stereo triangulation problem via two planes calibration procedure.

$$\Delta = \vec{P}_{12} - \vec{P}_{11}$$

where $\vec{\Delta}$ is a vector in world reference space defining the relationship between the projections of image pixels from the left and right cameras with respect to the world coordinate space. Again, from Duda and Hart [DH73] the location of the world point should lie at the head of the vector, $\vec{v}$ defined by

$$\vec{v} = \frac{a_0 \vec{u}_1 + (\Delta + b_0 \vec{u}_2)}{2} + \vec{P}_{11}, \quad (6.13)$$

where a_0 and b_0 are defined as

$$\begin{aligned} a_0 &= \frac{\vec{u}_1 \bullet \vec{\Delta} - (\vec{u}_1 \bullet \vec{u}_2)(\vec{u}_2 \bullet \Delta)}{1 - (\vec{u}_1 \bullet \vec{u}_2)^2}, \text{ and} \\ b_0 &= \frac{(\vec{u}_1 \bullet \vec{u}_2)(\vec{u}_1 \bullet \vec{\Delta}) - (\vec{u}_2 \bullet \Delta)}{1 - (\vec{u}_1 \bullet \vec{u}_2)^2}. \end{aligned} \quad (6.14)$$

At this point, it would be useful to gain some insight into the sensitivity of Equation 6.13 with respect to random position error in each of the vectors from the set $\mathcal{V} = \{\vec{P}_{11}, \vec{P}_{12}, \vec{P}_{21}, \vec{P}_{22}\}$.

6.4.1 Perturbation Analysis

Since the error in the each of the vectors from the set $\mathcal{V}$ was modeled and bounded in the monocular analysis, these estimates can be used to measure the propagation of residual error from the calibration phase to the final three-dimensional reconstruction. In fact, the norms on the residual error matrices cataloged in Table 6.6 will serve as the boundaries for the error analysis. As a result of the assertion that residual error is bounded by the estimates in the table, it can accordingly be stated that the error in the final three-dimensional reconstruction is bounded if no singularities are present in either of Equations 6.13 or 6.14. By inspection of these equations, it becomes clear that a singularity is achieved when the unit vectors $\vec{u}_1$ and $\vec{u}_2$ have the same orientation. In stereo vision, this corresponds to the case where a stereo camera pair is observing a point from an infinite viewing distance. Clearly, this case cannot occur in a real situation; however, it not hard to extrapolate from these equations that as the position of these unit vectors is perturbed some significant error could be introduced.

Sensitivity was measured in two ways. First, it was desired to determine how the stereo reconstruction algorithm performs under increasing distance from the object points. As revealed by the denominator of Equation 6.14, one would expect the residual error measured in the reconstructed scene points to experience a nonlinear increase as the distance to scene points is increased. Figures 6.5 and 6.7 present the magnitude of the percent absolute residual error in the stereoscopic reconstruction of

calibration targets at a distance of 45.72 cm. The target grid was uniformly distributed on the imaging plane to ensure global sensor coverage. For convenience, the graphs are arranged in pairs where an index indicates from which image plane half the error was measured. The figures indicate that for this standoff distance, maximum expected error falls around 1.1% regardless of image plane locations. A second measurement was included to synthesize a working boundary for this stereo system. A global quality criterion of 3% maximum residual error was used to define the valid region of operation. Figures 6.6 and 6.8 display the residual error for standoff distance of 124.46 cm. Analysis of data in these figures confirms that our quality criterion has not been violated. Using the stereo equations beyond this distance would incur an nonlinear decrease in their ability to accurately recover object points that would quickly destroy its usefulness. Operational distances could be easily extended through the use of lens systems with larger focal lengths; however, a corresponding decrease in the field of view would be experienced.

Second, each of the vectors from $\mathcal{V}$ was corrupted with uniformly distributed noise on the interval $[-\epsilon, \epsilon]$, where ϵ is one of the estimates from Table 6.6. By definition each of these vectors is constrained to move within the boundary of the calibration plane. This means that each vector from $\mathcal{V}$ will be allowed to move within a circle of radius ϵ on the plane that its calibration geometry defines. This operation forms a new set of disturbed position vectors denoted as $\mathcal{V}'$ where one of corrupted vectors from $\mathcal{V}'$ together with three complementary vectors from $\mathcal{V}$ are input into Equations 6.13 and 6.14. This process is repeated on the interval at increments of $\frac{\epsilon}{60.0}$, and the resultant error surface is plotted for each $\mathcal{V}$ and $\mathcal{V}'$ combination in order that some sort of conclusions might be drawn on the response of the triangulation algorithm to uncertainty in intermediate

points.

Figures 6.9 and 6.11 display the effects of perturbation on front calibration plane position vectors in terms of absolute position error. From this figure pair, several interesting conclusions can be drawn. First, it is observed that variation in Y -axis direction (perpendicular to the baseline of the camera systems) provokes a slow varying response in the error function when compared to the X -axis disturbance. Also a mirror symmetry exists between the two plots. Both of these phenomena can be described in terms of the stereo triangulation equations. Any disturbance in the X direction corresponds to a direct manipulation of one of the vectors used in the denominator of Equation 6.14. Depending on random error and equation nonlinearity, the reconstructed calibration point will actually be located in front or behind the actual calibration plane. Thus, perturbations that decrease the angle between these vectors will also decrease the amount of error realized in the reconstruction. correspondingly, perturbations that increase the angle will experience a rapid nonlinear increase in measured error. This relationship is not in general true for points not lying on the calibration plane, but it does provide a useful description of how the nonlinearity in the stereo equations amplifies the uncertainty from the position vectors from $\mathcal{V}$. Mirror symmetry follows as a direct consequence of the geometrical relationship between the triangulation vectors $\vec{u}_1$ and $\vec{u}_2$ and their dot product.

A disturbance in $\vec{u}_1$ that decreases the angle between them can be realized by an displacement of $\vec{u}_2$ in the opposite direction. Figures 6.10 and 6.12 contain similar results for the back calibration plane. One important point that must be conveyed depends on the understanding that these error functions are point functions. This means that the shape will vary according to their locations in the imaging plane and the random

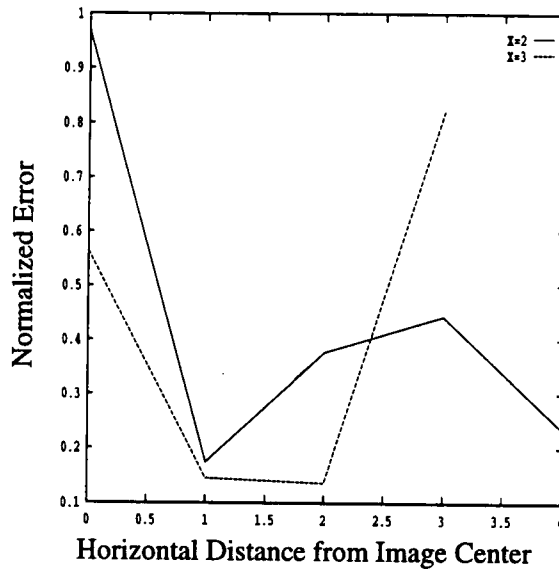


Figure 6.5: **Percent Residual Point Error:** Error measurements are obtained from the application of calibration points the stereo equations. Measurements are displayed for lower half of the image plane at a standoff of 45.72 cm.

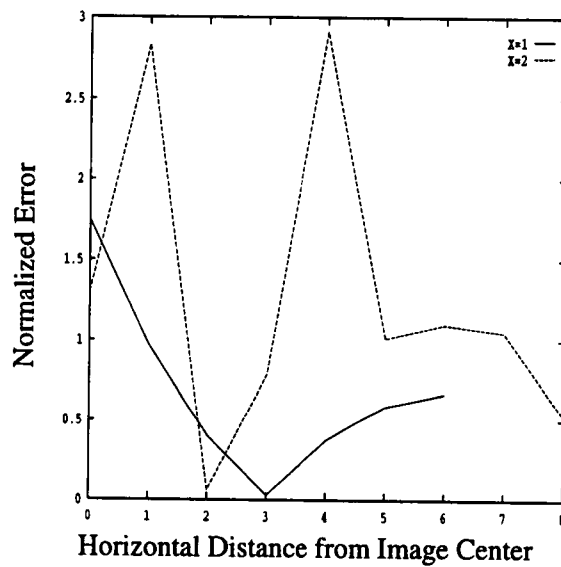


Figure 6.6: **Percent Residual Point Error:** Measurements taken from lower half of the image plane at a standoff of 122.5cm.

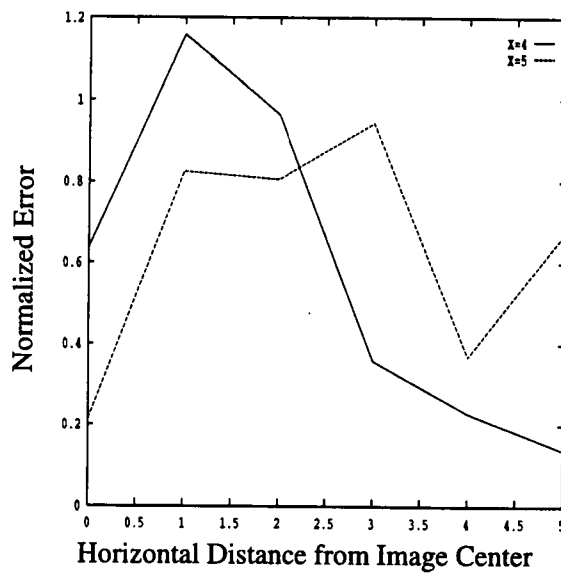


Figure 6.7: **Percent Residual Point Error:** Upper half of imaging plane at 45.72cm.

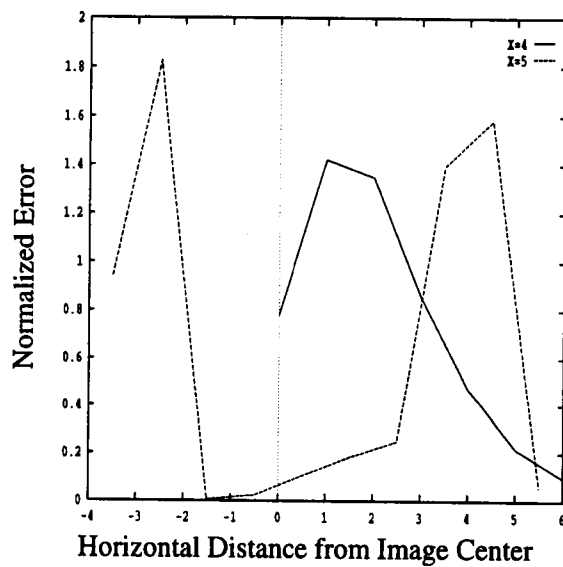


Figure 6.8: **Percent Residual Point Error:** Upper half of imaging plane at 122.5 cm.

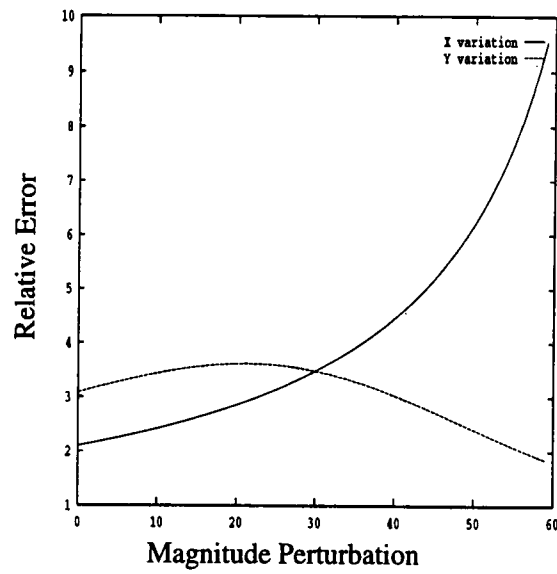


Figure 6.9: **Sensitivity of Stereo Equations:** Uniform Perturbations in $\tilde{P}_{11}$ derived from normal analysis of monocular residual error.

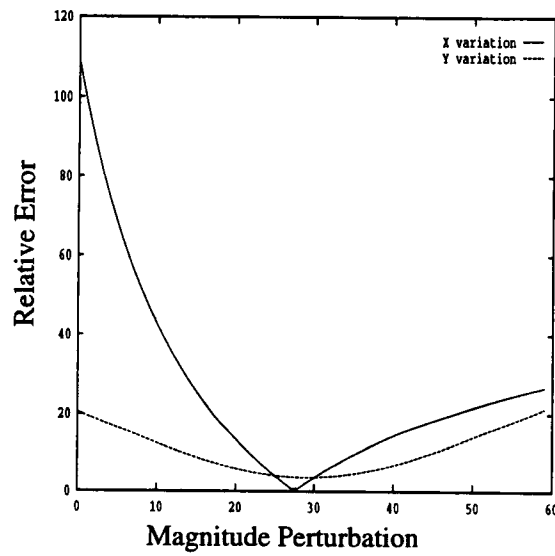


Figure 6.10: **Sensitivity of Stereo Equations:** Uniform Perturbations in $\tilde{P}_{12}$ derived from normal analysis of monocular residual error.

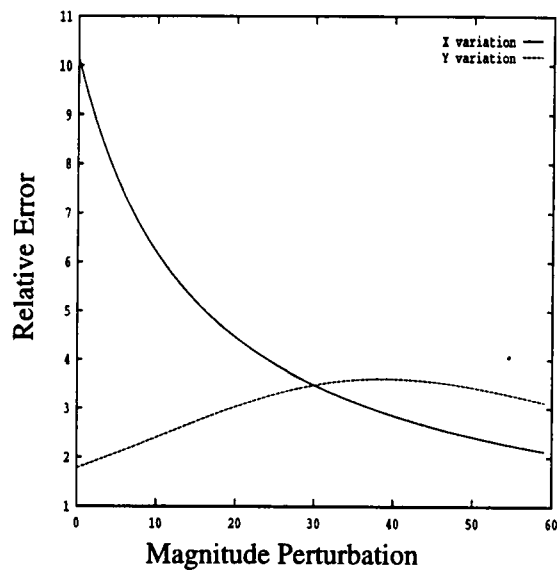


Figure 6.11: **Sensitivity of Stereo Equations:** Uniform Perturbations in $\vec{P}_{21}$ derived from normal analysis of monocular residual error.

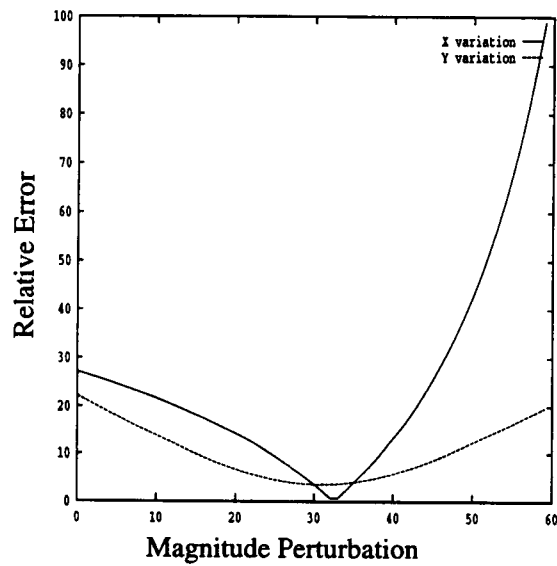


Figure 6.12: **Sensitivity of Stereo Equations:** Uniform Perturbations in $\vec{P}_{22}$ derived from normal analysis of monocular residual error.

error in the stereo reconstruction; however, the basic relationships that indicate that random displacement in the direction of the camera system baseline vector will realize large nonlinear shifts in the reconstruction error while displacement perpendicular to the baseline vector will not be as severe.

6.5 Conclusions

An in-depth mathematical analysis of the calibration mechanism was performed. Depending on the order of interpolation, the transformative process was found to be numerically stable over the operating parameters via a singular value decomposition and perturbation analysis. It was discovered that if the number of calibration target permits, singular value decomposition provides the most stable and an accurate solution. The existence of a direct relationship between the condition numbers of the Moore-Penrose pseudo inverse and those procured directly from the interpolation matrix proved this. In cases where numerical resolution or other constraints requires polynomials that a specific data representation format cannot handle numerically, a method for removing troublesome interpolation terms before their ill effects can propagate was presented and found to be successful.

Analysis of the residual error realized in triangulation of calibration targets and the consequential reconstruction of the calibration planes reveals the nonlinear relationship between the two-dimensional residual error bounds synthesized in the monocular analysis and the propagation of this error through the stereo equations. It found that uncertainty along the camera baseline vector is numerically more significant than displacement perpendicular to the baseline vector. Examination of reconstructed cali-

bration planes revealed that the maximum uncertainty in point location of the stereo system with standoff distances less than 122.5 cm would be less than 3% as a linear function of distance. Beyond this range additional optics would be required to improve performance to an acceptable level.

CHAPTER 7

Representations for Integrated Active Stereo

Tarr and Black rise to the defense of Marr's paradigm with a computational and evolutionary argument along with flanking counter-attack in a proud military fashion. First they assure us that "It is *flexibility* that distinguishes this reconstructive approach from the purposive approach" [TB94a]. They claim that it is the goal of vision to build symbolic scene descriptions and use "cognitive" operations on these models to interact with the world. Evolutionary justification for this hypothesis lies in their extension of a principle of natural selection of human languages to vision. Computational justification relies more on finger-pointing than fact and can be summed by their statement, "The reconstructive approach is no-more mired in the ill-posed nature of visual perception than is the purposive approach" [TB94a]. They begin their criticism of the active paradigm with misunderstanding [Alo94] when they (Tarr and Black) reveal "the goal of an active vision system is to reformulate traditionally static vision systems in terms of dynamic exploration of the environment." Next, they complement their mistake by ensuring us that the introduction of active constraints into previously recorded data is a promising technique for the development of *robust* applications. Their criticism of the active paradigm is primarily concentrated at the purposive level and is lacking in sufficient justification. They conclude by saying, "...we believe it is better suited for understanding and mimicking the overall visual behavior of frogs rather than humans." Their final response [TB94b] is essentially a concession that both ideal *Reconstruction*

and *Purposivism* are elusive and worthy goals.

7.1 Representational Philosophies

Tarr and Black, in their defense of reconstruction, do raise several excellent questions that remain unanswered. Some questions are biological in nature and center around the evolutionary pressures to evolve one type of vision over another. While these questions are intriguing and certainly merit study, they are somewhat misplaced in an analysis of practical vision. Indeed, solution to those problems may be ill-posed as well! With all of this aside, concentration can be set on the main question their text raises. **What types of representation need to be internalized in vision?** Should we seek to build vision with abstract representations that are suited for general-purpose tasks such as reasoning. Or should we use task-level representations to simplify and strengthen our algorithms? Perhaps, representations should be foregone entirely in the sense of Brooks [Bro91] and synergistic and emergent properties sought. If these questions could be answered in the optimal sense at the systems level, objective judgments on the merits of representation in any form would be verifiable.

7.1.1 Challengers with Purpose: The Purposivist

Aloimonos is the most ardent of the purposivists and his views represent its extremum. He claims that purposive vision is capable of producing *flexible* and *intelligent* vision, and may, in fact, offer insight into the operation of the human brain. He has also implied that purposive vision is the natural extension of the active paradigm. He further states [Alo94],

the fundamental problem of a vision system is determining *what information* from the image should be used and *what representation* of it needs to be built so that the relationship of the system to the world can best be implemented.

From this statement it can be seen that Aloimonos does not advocate foregoing internal representation as Brooks; in fact, he is saying the opposite. Vision must use optimal representation as a function of the task. Aloimonos' views with respect to recovery are inflexible. He retorts, "I can now tell you that vision has nothing do with recovery; vision is a recognition process". [Alo94]. He qualifies this with four characteristics of vision. Vision is

- Associative
- Interpretive
- Directed
- Selective

However, what is not clear from his statement is to what "vision" he is referring. It is not the precise recovery of depth used in photogrammetry, nor is it the state of computer vision today. He is describing characteristics of vision in general. The same general vision that he has declared a "chimera". Human or general vision may indeed embody these four qualities; however, computer vision most definitely does not. Computer vision is what we make of it; it can be as quantitative or as qualitative as it needs be. In this respect Aloimonos has lost perspective. Blind statements of the merits of subjective qualities of visual system are not productive. Someone once said, "The proof is in the

pudding". And before any ideas from either camp should be adopted, an objective, quantitative analysis of vision systems using both methodologies must be pursued.

7.1.2 PeaceMakers: The Realists

Tsotsos [Tso94] forces us the fundamentals of the issue with his analysis of the tractability of each of the paradigms thus presented. He relates that the Marr paradigm, the active and purposive paradigm, as well as the methodology of Brooks are NP-hard in their general formation [Tso90]. If we accept this as true, it comes as an astounding revelation. It implies that arguments supportive of the computational efficiency of one of these paradigms over another is meaningless. Computational comparisons must occur at the systems level before a meaningful quantification can be assigned. We may say that vision system A is more computationally efficient than vision system B, but cannot say that active/purposive/qualitative visual theory is framed in a more favorable computational structure. We must look elsewhere for our justification, but where?

Fischler focuses on what he believes is central to this debate, "... the extent to which it is possible to construct a meaningful representation of a scene without an understanding of the purposes for which the representation will be used" [Fis94]. From this, it is quite clear that he is supportive of using task-level knowledge in the design of visual systems; however, he criticizes the notion of a general or abstract representation for intelligent vision. In his mind, correspondence to the physical world is a fruitless basis for representation as it violates the very definition of a representation scheme - to make some knowledge explicit and eliminate some knowledge. In this context, a general representational scheme becomes impossible since there is no way to intelligently disregard some information and select other information without some knowledge of task. Sandini

and Grosso [SG94] hold the similar view that reconstruction is axiomatically purposive. In this framework, representation does make explicit use of task, and if vision systems perform a task then conversely they must use some representation. Advantage, according to both groups, is gained from the using the minimal representation dictated by the purposive approach.

Christensen and Madsen [CM94] view the active and Marr paradigms as complementary processes where basic research in vision is conducted via reconstruction which is combined with purposive level analysis into functional systems. Jain [Jai94] agrees when he asserts, "representations are central to any computational approach", and also says, "Every vision system ever developed was for a purpose and will be for a purpose". Brown [Bro94] concentrates on both finding areas of deficit and highlighting assets in both paradigms. Purposive vision, as he maintains, may discourage thinking about *General* vision which he has defined as human vision. He also claims that the behavioral context in which many purposivist present their work is exclusive and does not prove useful in any way. According to Brown, purposivism's main asset is its drive to intelligently integrate vision architectures. As for a reconstructionist approach he relates, "the reliable, world-linked, cognitively impenetrable status of reconstructed representations is one of their most attractive features". [Bro94] In saying this, he is telling us of the merit of reconstruction. Aggarwal counters with, "the seductive promise of reconstruction is that if we can just get more information from the external world and build it into our internal models, we then will be able to *easily* pick and choose the specific information we need."

7.1.3 What has been Learned

The idea of researching the usefulness of a representational scheme is directly motivated by its possible uses. As the principle thesis of this work stated: no single form of representation can be considered adequate or even feasible for all possible applications. Thus the goals of this work can be directed to finding efficient, flexible, and accurate representations for our task or application. In particular, the difference between representations for identification and localization tasks will be examined. In doing this, range sensing can be fully optimized for input into a perception or learning algorithm. Some of the benefits expected from this endeavor are

- Reduced storage costs,
- Increased stability and robustness in learning algorithms,
- Reduced complexity of vision programming,
- Improved efficiency in the task domain, and
- Direct mapping into an viable computational theory.

Any representational form for active stereo, or any task that seeks to make use of vision, should be designed to implement at least a partial set of the guidelines outlined in the previous list. Ideally, it should facilitate all of them; however, since no representation is capable of meeting all the needs of an active vision system. The baseline representation should be loosely structured and highly tunable for a particular task. Also, depending on the vision task it should be able, in its limit, to faithfully represent systems that require use of any of the basic representational schemes, *Recovery*, *Purposive*, or *Reactive*, with

minimal effort. Therefore, we look to a basic mathematical theory that provides the capabilities of looking at stereo imagery with the capabilities to simplify system-level vision.

7.2 Parametric Elliptical Fourier Representations

Any closed, continuous curve can be decomposed into a parametric representation using Fourier series coefficients. In systems with a single independent variable, Tolstov [Tol62, Goo68, OW83] relates that one dimensional signal with period 2π can be represented by

$$x(t) = a_0 + \sum_{k=1}^{\infty} (a_k \cos(kt) + b_k \sin(kt)). \quad (7.1)$$

Staib, et al. [SD92, KG82, LH87] show that any two dimensional plane curve can be decomposed in to a parametric Elliptical Fourier representation as

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} a_0 \\ c_0 \end{bmatrix} + \sum_{k=1}^{\infty} \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} \begin{bmatrix} \cos(kt) \\ \sin(kt) \end{bmatrix}. \quad (7.2)$$

In both equations 7.1 and 7.2 individual elements of the matrices are obtained from

$$\begin{bmatrix} a_0 \\ c_0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2\pi} \int_0^{2\pi} x(t) \cos(kt) dt \\ \frac{1}{2\pi} \int_0^{2\pi} y(t) \cos(kt) dt \end{bmatrix}, \quad (7.3)$$

and

$$\begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi} \int_0^{2\pi} x(t) \cos(kt) dt & \frac{1}{\pi} \int_0^{2\pi} x(t) \sin(kt) dt \\ \frac{1}{\pi} \int_0^{2\pi} y(t) \cos(kt) dt & \frac{1}{\pi} \int_0^{2\pi} y(t) \sin(kt) dt \end{bmatrix}. \quad (7.4)$$

Using a truncated version of the series in Equation 7.2 and writing in matrix notation, any closed, continuous, two-dimensional plane curve can be approximated by

$$\vec{B}(t) = \vec{C}_o + \sum_{k=1}^N \mathbf{C}_k \vec{T}_k(t). \quad (7.5)$$

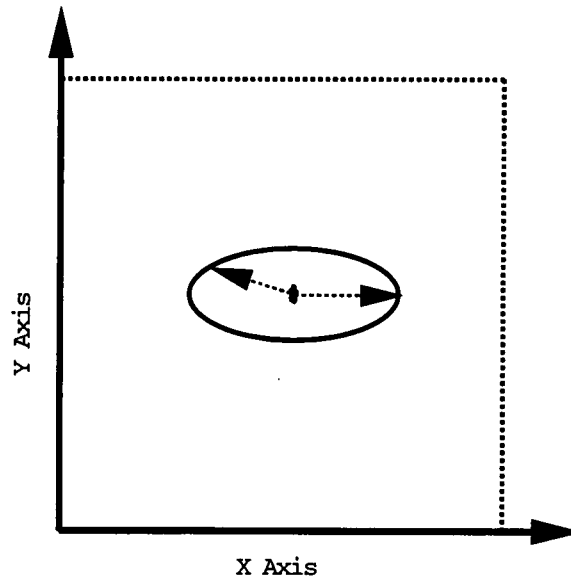


Figure 7.1: **A Perfect Ellipse:** Parameterization generates two one dimensional data sets by scanning the x and y coordinates into individual vectors.

In this form, $\vec{C}_o$ represents the origin of the parametric decomposition which is computed from equation 7.3 and the 4×4 matrix $\mathbf{C}_k$ is derived from equation 7.4. The sinusoidal basis vector, $\vec{T}_k$ is shown in equation 7.2. The order of the approximation, N , controls the amount of high frequency components used in the representation. For instance, circles and ellipses can be modeled using an order of one, while points are of order zero. (Note: all curves referred to as plane curves are assumed to be continuous and closed unless specifically mentioned in the text)

Figure 7.1 shows a perfect ellipse in two dimensions. A parameterization is obtained by radially scanning from the center of mass and generating two line profiles as shown in Figure 7.2. From here, a Parametric Elliptical Fourier Representation (PEFR) is obtained using equation 7.5 with the user selecting the desired order depending on the characteristics of the input data. In this case, an order of one should be sufficient to

model the characteristics of the input data. Table 7.1 contains the PEFR coefficients

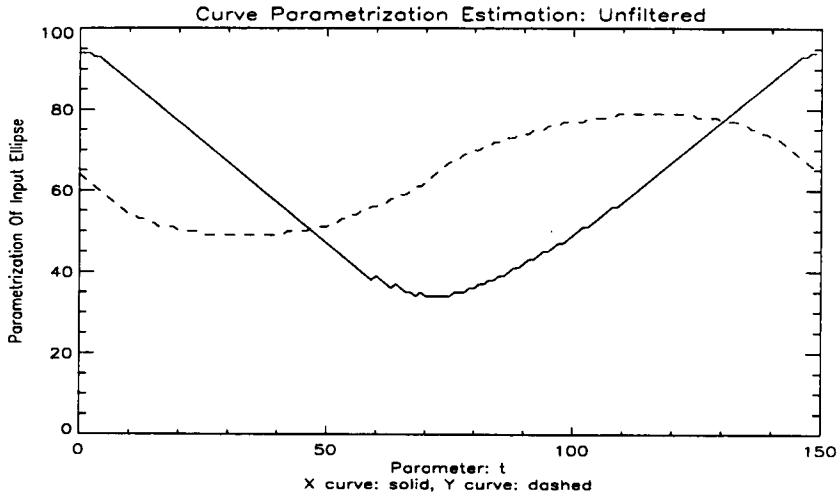


Figure 7.2: **Parameterization of a Perfect Ellipse:** Discretization of the input curves causes high frequency noise to be imposed onto the underlying fundamental sinusoids.

up to third order for the perfect ellipse(see figure 7.1) input to the system. In the table, discretization error has added noise onto the individual Fourier components; however, the basic structure of the ellipse is recovered to within 4% for curves of order one and gradually improves with more coefficients.

In addition to the fundamental matrix equation for PEFRs(shown in Equation 7.5), a weighted curve representation that penalizes selected frequency components is proposed. Thus it will enable the user to design at representation with band rejection properties. This is extremely important when the frequency components of template models are well understood. The frequency weighted PEFR is related as

$$\vec{B}(t) = \vec{C}_o + \sum_{k=1}^N \alpha_k^k \mathbf{C}_k \vec{T}_k(t), \quad (7.6)$$

where α_k is a weighting scalar on $[0, 1]$

Table 7.1: Estimation of the Input Curve: with varying orders

Order	$N = 1$	$N = 2$	$N = 3$	Original
a_0	62.83	62.83	62.83	64.0
c_0	64.11	64.11	64.11	64.0
a_1	28.79	28.79	28.79	30.0
a_2	0.0	0.48	0.48	0.0
a_3	0.0	0.0	0.0.37	0.0
c_1	15.88	15.88	15.88	15.0
c_2	0	-0.1	-0.1	0.0
c_3	0	0.0	-0.14	0.0

7.2.1 PEFR's as a Representation for Active Stereo

In general, a three level hierarchy of stereo matching must be maintained. *Points*, *Regions/Area*, and *Boundaries* [Tha93, NP83, Gri93] are typical representations used to match image structures in a corresponding stereo pair. As will be shown, PEFR's will provide a capability to synthesize all three structures, while fulfilling at least one of our representation requirements, e.g. reduced storage costs.

Points PEFRs

Point-based operation is linked to the reconstructionist paradigm of Marr [MH80, Mar82]. Matching and correspondence algorithms for image points, or pixels are well documented [AC94, ABG89, SS90]. Since the body of literature for point representations is exhaustive, a simple relation of how degenerate PEFRs can simulate points is offered.

From equation 7.5, given any plane curve, a zero order approximation to that curve is the center of mass. In space, this is registered as a single point, not necessarily on the boundary of the curve. Therefore, using PEFRs as points is a simple matter of assigning the location of the point to the vector, $\vec{C}_0$. The equation for this construct is

$$\vec{B}(t) = \vec{C}_0, \quad (7.7)$$

which is clearly independent of the parameter t . A matrix, or image, is constructed similarly.

Plane Curve (Boundary) PEFRs

In order to simplify operations on plane curves, we look to a new centered representation that lends itself well to radially sampled curves. Therefore, given a 1×2 vector $\vec{B}$, a centered version is given by

$$\hat{\vec{B}}(t) = \vec{B}(t) - \vec{C}_0 = \sum_{k=1}^N \mathbf{C}_k \vec{T}_k(t), \quad (7.8)$$

where the right side of the equation is trivially obtained from equation 7.5. This means that plane curves can be generated and operated on at the origin and then translated to their actual location in space.

The next point that needs to be addressed is the resolution or coarseness of the representation. Given that we are interested in plane curves of a digital nature, equation 7.8 and a set of Fourier coefficients, $\mathbf{C}_k$, that is scanned radially in $\Delta\theta$ uniform increments along the interval $[0, 2\pi]$ will produce an $(M+1) \times 2$ curve template matrix, called $\mathbf{S}$. The rows of matrix $\mathbf{T}$ each specify a unique point along the curve boundary, with the exception of the last point which is a replicate of the initial point to ensure

continuity. The operation of creating S is encapsulated by

$$S_i^T(i\Delta\theta) = \hat{B}(i\Delta\theta) + \tilde{C}_0, \quad (7.9)$$

where the parameter $t = \Delta\theta = \frac{2\pi}{M}$ and i is the specification of the row of matrix S . Now, a centered, variable resolution template of a plane curve is available. This representation will form the analytical basis for the rest of the chapter.

Using a variable resolution representation will facilitate computational efficiency and matrix storage requirements. However, one must be wary of sampling issues in the reconstruction of PEFRs from the original figure. In fact, Shannon's sampling theory relates that in order to generate an adequate representation our parametric curves must be sampled at two times the rate of the highest frequency present in the curve in order to avoid aliasing. Figure 7.3 shows two possible forms of aliasing that will be encountered in the use of PEFRs. The first form is a direct consequence of Shannon's theorem and it is manifested through inadequate sampling of the parameterization. The rows [2 – 6] of the figure matrix are PEFRs with varying sampling rates and constant order. The original image was a simple square shown as the very first figure in the diagram. As shown in the first few columns, the aliasing is severe enough to form degenerate shapes. Next, a phenomenon caused by inadequate use of high frequency terms in the PEFR is shown in rows two and three respectively. In these two rows, the PEFR cannot mimic the structure of the input curve regardless of the sampling rate. This is corrected for in the last rows of the matrix where the PEFR order is somewhat higher. Thus, two hard constraints must always be satisfied when using PEFRs as a representation in any system. They are:

- Sampling of the parameterization must satisfy Shannon's sampling theorem.

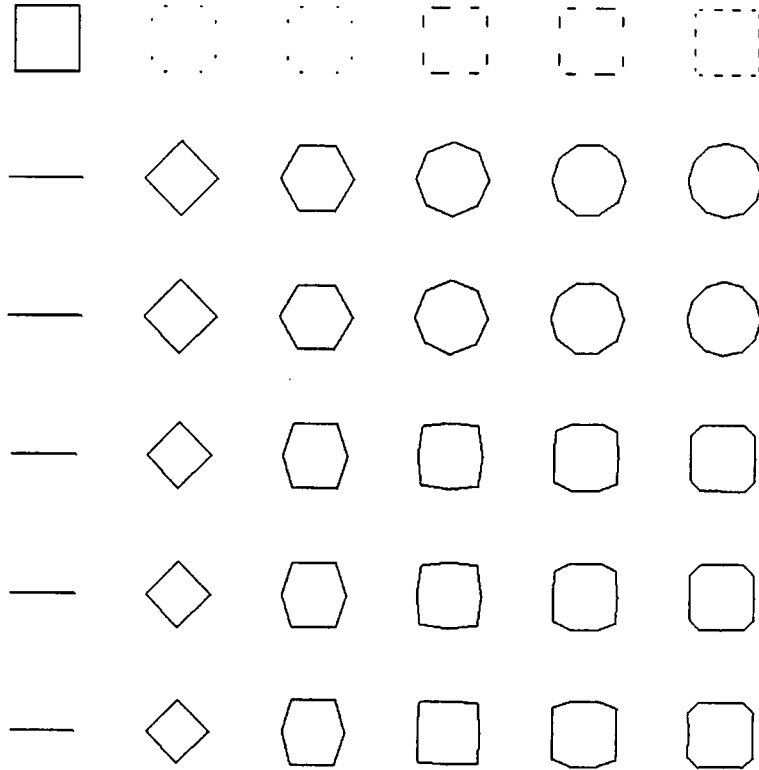


Figure 7.3: **Sampling vs Order Grid:** Rows [2 – 6] of the figure matrix have constant order [1 – 5]. Columns of the figure matrix have constant sampling rates [2, 4, ...10, 12]. The first row contains the original image followed by the intersection of it with the ideal PEFR with 500 samples and orders [1 – 5], respectively.

- PEFR order must be sufficiently matched to the input data characteristics or degenerate structures will emerge.

Region PEFRs

A homogeneous region of space is implicitly described by discrete curve map, S where the resolution of the region is controlled in equation 7.9. Given this curve description, it is possible to design a function $f(S)$ called a filling function that produces a uniform

region bounded by the curve S . This function is formally defined as

$$\mathbf{R} = f(\vec{F}, k) = \begin{cases} k & \text{where, } x(a), y(b) \text{ on the interior of } S \\ 0 & \text{otherwise} \end{cases} \quad (7.10)$$

For regions the operations are derived from the filling function on a centered plane curve, however these operations are not limited to centered functions due to their association with basic set theory. The first operator is the familiar union operation. This essentially constructs a composite PEFR that is the union of N different PEFRs. It is defined as

$$\tilde{\mathbf{R}} = \bigcap_{i=1}^N \mathbf{R}_i. \quad (7.11)$$

Likewise, an intersection for N regions is defined as

$$\tilde{\mathbf{R}} = \bigcup_{i=1}^N \mathbf{R}_i. \quad (7.12)$$

7.2.2 Fundamental PEFR Operators

A filling operation has already been defined in the previous section that augments a plane curve by filling of its interior with a constant scalar value. This has many uses in the analysis of homogeneous regions that can be thus simulated. In addition to this functional operator, several additional operations are implied since the basis of PEFR curve theory lies in the Fourier series. (Note: Unless specifically mentioned all basic operations are plane curve specific. Operators on regions must first be applied to the base plane curve and then the filling function can be used to generate a valid region.) The first operator comes as a direct consequence of the Fourier series property of addition. Which relates that the Fourier series of a sum is the sum of the Fourier series. This directly lends itself to define a blending function for a set of N centered

plane curves. Thus, a centered N -blend is defined as

$$\hat{\vec{B}} = \sum_{i=1}^N \alpha_i \hat{\vec{B}}_i, \text{ where } \alpha_1 + \cdots + \alpha_N = 1.0. \quad (7.13)$$

Clearly, an average of N curves can be obtained by the constraint of equivalent coefficients. However, this structure will allow the use of prior knowledge as a weighting coefficient. Scaling is another property that falls directly from Fourier series properties. The scaling property is thus related as

$$\alpha \hat{\vec{B}}(t) = \sum_{k=1}^N \alpha \mathbf{C}_k \vec{T}_k(t) \quad (7.14)$$

Differentiation is also expressed implicitly in this representation. If equation 7.8 is rearranged such that the coefficient matrix is separated in two column vectors such that

$$\hat{\vec{B}}(t) = \sum_{k=1}^N [\mathbf{C}_k^1 \mathbf{C}_k^2] \vec{T}_k(t), \quad (7.15)$$

where

$$\mathbf{C} = [\mathbf{C}_k^1 \mathbf{C}_k^2], \quad (7.16)$$

then

$$\hat{\vec{B}}'(t) \approx \sum_{k=1}^N k [\mathbf{C}_k^2 [-\mathbf{C}_k^1]] \vec{T}_k(t). \quad (7.17)$$

Likewise, an integration of the curve is available as a consequence of Fourier series properties. The formal expression is

$$\int_a^b \vec{B}(t) dt = \frac{\vec{C}_o}{2.0} (b - a) + \sum_{k=1}^N [\mathbf{C}_k^1 [-\mathbf{C}_k^2]] [\vec{T}_k'(t)]_b^a. \quad (7.18)$$

Notice this representation provides a number of convenient operators to deploy in system design; however, it has the disadvantage of making the curve length difficult to recover using simple operations on Fourier coefficients.

7.3 Conclusions

A basic theory of parametric curve representation was outlined. This theory relies heavily on the Fourier series for many of its fundamental advantages. In addition, a group of operations that can be used directly on these representations was outlined. Finally, a representation capable of multiple resolution analysis was derived. Using this basic tool combined with the Fourier series properties offers many advantages. In terms of the guidelines for adequate representation that was outlined in the beginning of the chapter, three main advantages, computation efficiency, reduced storage costs, and direct mapping into computational theory were realized.

7.3.1 Computational Efficiency

The use of a variable resolution PEFR can be directly quantified through the use of equation 7.5 and basic knowledge of digital imagery. Two tiers of computational advantage can be identified in the PEFR schema. The first is inherited directly from the properties of a Fourier series. All of the aforementioned operations have been reduced from operations on plane curves to operations on Fourier series. Thus, many complex systems can be designed and built to operate on the internal representation rather than the original image resulting in a computational advantage. For an excellent discussion on the computational advantages of the Fourier series see Tolstov [Tol62].

In many cases, the internal representation must be converted into a form that is compatible with image-based operations. In this case, the computational advantages associated with the Fourier series are forfeited in favor of those in equation 7.9. For example, the operation of template matching which is a common operation consists of a

multiplication followed by an integration to give the total response of the template to the image. This operation is typically formed by direct multiplication of the template image onto the original image. This results in $2I^2$ additions and multiplications, for square images, where I is the dimension of the image and is typically quite large. A more efficient concept of template matching is related in [SD92], where the dimensionality of the operation is reduced to a function of the curve length given in equation 7.18. Here the total cost for operations is $2.0 * L$, where L is the length of the curve in pixels. Variable resolution PEFRs further reduce this computation by the M/L . So the total number of operations is $2.0 * M$ where $M \ll L$.

Table 7.2 contains hard numbers for the three different template matching mechanisms previously outlined. In the table, the number of operations for full image template matching grows with the square of image size. It is independent of the template geometry. In full curve template matching the discretized curve length forms the basis of the number of operations. In this example an ellipse that grows in proportion to image size was used; accordingly, the number of operations grows with the baseline curve length. For mutli-resolution PEFRs a constant sampling angle of 6° was employed, giving a constant cost for all imaging sizes.

7.3.2 Storage Requirements

The storage requirements of a frequency weighted PEFR are a simple function of the order of the PEFR and the internal representation used in the digital computer. The unit size is related by

$$S_B = (5.0N + 2.0) * \text{sizeof}(ir), \quad (7.19)$$

Table 7.2: **Operation Cost Comparison:** for three different types of template matching schemas

Image Size	Full Image	Full Curve	PEFR
64×64	8192	75	60
128×128	32768	150	60
256×256	131072	300	60
512×512	524288	600	60

where ir is the internal representation used in the system and the “sizeof” operator gives the number of bytes for that input N is the order of the PEFR. The size of the PEFR image vector from equation 7.9 is given by

$$S_S = 2(M + 1) * \text{sizeof}(ir) \quad (7.20)$$

Table 7.3 contains a comparison of the storage costs associated with PEFRs of varying resolution and order. The units in the table are “bytes” and represent the total combined storage cost of a PEFR with order N and an image vector derived from that PEFR with samples M .

7.3.3 Direct Mapping in Computation Theory

The issue of mapping into a viable computational theory is a major issue in the use of any internal representation. In particular, it is desirable to make certain major fruits of that theory available to the vision programmer in a trivial manner. From section 2 of this chapter, certain efforts were made to realize part of this goal. A fundamental set of operators was presented that facilitates simple mathematical operations on the chosen representation. However, the major benefit of using PEFRs was only hinted at

Table 7.3: **Storage Cost Comparison:** with varying resolutions and orders

Order	Samples					
	10	20	30	40	50	60
1	29	49	69	89	109	129
2	34	54	74	94	114	134
3	39	59	79	99	119	139
4	44	64	84	104	124	144
5	49	69	89	109	129	149
6	54	74	94	114	134	154

in the previous subsection. It centers on the use of template and model matching to perform tasks such as stereo matching, object identification, and regional analysis. In using PEFs to form templates from a binocular image pair and/or individual images, Maximum A Priori (MAP) estimates can be obtained in an explicit manner using a PEF and an input image set. While the explanation of this is left in detail to the remaining chapter, it allows the inclusion of a large, well-developed body of optimization theory to be included directly into vision design. Coupling this with explicit MAP estimates using multi-resolution PEFs, powerful visual systems can be designed using the baseline architecture presented in the following chapter. This architecture will bind local representation, computational theory, and task-domain knowledge into an algorithm well-suited to handle difficult problems in vision.

CHAPTER 8

Architecture for Purposive System Design

Many designers use *sequential block algorithm* design as a simple road map to implement a visual system. In this design methodology, an algorithm is decomposed into a network of sequentially connected sub-components. These sub-components, or modules, implement some specific computational function, e.g., edge-detection, thresholding, etc. By connection of the output of one module to the input of another module, complex algorithms can be constructed in a simplified framework. In a similar manner, individual algorithms can be viewed as functional blocks and assembled into a complex system. The primary advantage of this type of linear decomposition is the speed of implementation and the warehousing of functionality into a conceptually simple representation. While this class of methodology offers many advantages to the designer, it can produce inadequate results for highly dynamic vision applications due to the propagation of component error from block to block. In this type of degeneration, a sequence of individually robust modules can lead to global algorithmic instability. In this work an alternative to the sequential block design is presented.

A Bayesian Algorithmic Network Design, or **BAND**, is proposed. It uses a system of estimation algorithms together with their Bayesian priors and evaluation functions embedded into a matrix that describes the network. Given this design matrix, the question is posed, "Which sub-component of the Bayesian network has offered an optimal result in light of the given information?". The richness of this structure is provided

through the flexibility of optimization techniques and the inclusion of task-level knowledge that can tailor the BAND network to a particular application. For instance, a static BAND network can be constructed to provide a class of edge-detectors for two-dimensional medical images. Each individual edge detector may have been constructed under the sequential block algorithm, or SBA, paradigm. The BAND network can then be optimized to adjust the Bayesian priors to favor detectors that return single pixel edges, or weight the priors based on system timing information for time critical applications.

This analysis begins with a break-down of the SBA design philosophy and a demonstration of how component error propagation can lead to algorithm failure. From here the BAND theory is developed and presented in the context of system design. Next, simulated results are discussed. Chapter 11 offers an implementation of a BAND network for the extraction of cranial structures from positron emission tomography (PET) images.

8.1 Sequential Block Algorithm Design

Figure 8.1 conveys the traditional mode of vision algorithm construction. A task is decomposed into a sequence of N processes that operate in a sequential manner; that is, the input to process i is dependent on the output from process $i-1$. (Note: More complex algorithm design is possible in this paradigm using loops, feedback, etc.; however, for the purpose of simplifying the analysis of these structures, it has been assumed that more complex operations are encapsulated within the function of a particular process.) With each process, P_i , a value called the probability of success can be assigned to that

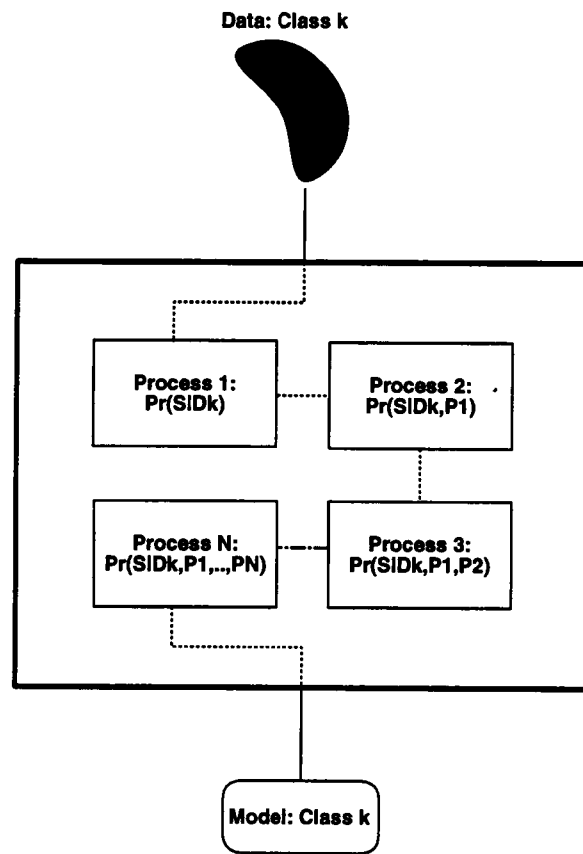


Figure 8.1: **Sequential Block Algorithm:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.

process. This value may or may not have a computational basis; however, it is viewed as the chance that the process, P_i , was successful given the input sequence. To further simplify the mathematics, it is assumed that the output of each process is independent of the outputs from the previous processes. This system could be extended using a Markov state model where the probability of P_i is dependent on the success of the previous states; however, the Bayesian model will provide an upper bound measurement on the success of SBA design. This quantification will provide the best baseline for our comparison

Table 8.1: **Sequential Block Algorithm (4 process algorithm)**: Showing how a sequence of relatively robust processes can lead to systems with low reliability

Process	$Pr(S_k^0)$	$Pr(S_k^1)$	$Pr(S_k^2)$	$Pr(S_k^3)$
P_1	0.97	0.90	0.80	0.85
P_2	0.92	0.85	0.75	0.53
P_3	0.95	0.95	0.86	0.82
P_4	0.99	0.99	0.69	0.65
$Pr(S)$	0.84	0.72	0.36	0.24

with BAND systems.

Elementary probability theory states the the probability of the union of N independent events is the same as the multiplication of their individual probabilities. Thus, given the N processes from Figure 8.1, the probability of success of the algorithm is computed as

$$Pr(S_k) = \prod_{i=1}^N Pr(S_i | D_{k,i-1}, P_1, \dots, P_{i-1}). \quad (8.1)$$

And in light of the independence assumption from the previous paragraph, is equivalent to

$$Pr(S_k) = \prod_{i=1}^N Pr(S_i | D_{k,i-1}). \quad (8.2)$$

Equation 8.2 illustrates that given the state of input data after the preceding step, $D_{k,i-1}$, and a vector that measures the confidence of a component process, P_i , the probability of the success of the system can be derived. Table 8.1 displays the confidence in 4 different SBA designs with strong to medium confidence in the individual component blocks. Even in blocks where confidence is greater than ninety percent for each block, the overall system confidence erodes rapidly. In addition to this drawback, SBA algorithms

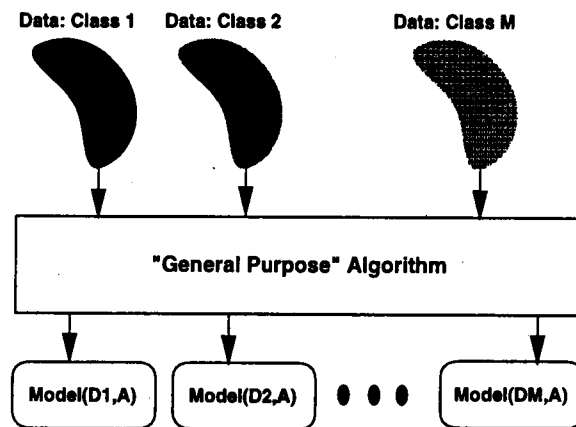


Figure 8.2: A “General” Algorithm: A simplistic model of algorithm design where an algorithm has been tuned for a particular class of input data. If it is used in the classification of a different class of data, performance will suffer.

usually suffer an additional problem that results from the design of “general” vision modules.

8.1.1 General Algorithms and SBA

Almost from the beginnings of computer vision research, it was the desire of researchers to build general vision modules. In early computer vision thinking, these modules would perform in a robust manner across all combinations of input data. Figure 8.2 shows a system where a single algorithm is responsible for deriving models from any class of data that is presented. As research into practical computer vision progressed it gradually became evident that this type of generic system is only a theoretical construct and could not possibly have any mapping into a realizable system [Alo90b, Alo89, Alo90a]. In contemporary computer vision, an algorithm must be designed for a particular visual task. In other words, we build algorithms as a function tuned to the input data, and the algorithm constraints must come from requirements of a specific task.

Table 8.2: **General Algorithm (1 process, 4 data classes)**: An algorithm that is designed for a particular data set may suffer instabilities when presented with dissimilar types of input

Sequence	D_0	D_1	D_2	D_3
$Pr(S D_k, 1)$	0.97	0.84	0.78	0.82
$Pr(S D_k, 2)$	0.95	0.93	0.64	0.45
$Pr(S D_k, 3)$	0.95	0.96	0.59	0.53
$Pr(S D_k, 4)$	0.99	0.90	0.61	0.42
$Pr(S D_k)$	0.96	0.91	0.65	0.55

Table 8.2 displays the results of 4 different trials of an algorithm designed to delineate structures in the brain. The primary algorithm was trained on input from class 0. From the table, the performance of this algorithm degenerates when presented with data from increasingly dissimilar data sets. The data set types were as follows: (0) FDG with 3D reconstruction, (1) FDG with 2D reconstruction, (2) 0-15 Water with 3D reconstruction, and (3) 0-15 Water with 2D reconstruction. Clearly the algorithm was designed for primary use with FDG systems.

8.1.2 Looking for a Better Framework

Based on the these simple preliminary studies, a framework for purposive vision has no place in linearly-sequenced, general vision modules. Linear sequencing can result in poor system performance and general vision design is impractical at best. Theories of purposive vision relate that each algorithm must be designed with a task in mind, i.e., each algorithm is tuned to a particular data class. Again, it is equally impractical

to design a set of algorithms for each set of related data in problems that may require hundreds of differing data streams. An answer lies somewhere in the middle. Robust systems can be designed to reference more than one data class with less than a cardinal number of algorithms to do it. By taking advantage of the full computational power of a minimal algorithm set and exploiting as many high-level purposive constraints as possible, robust task-level vision will emerge.

8.2 Parallel Block Algorithm Design

In selecting a design methodology for a purposive active vision system that extends the capability of the traditional SBA design, it is desirable to encode into this architecture two useful system design strategies that will facilitate its utility. They are: (1) implicit parallelism, and (2) dynamically reconfiguration [FF95]. It is also recognized that to provide a fundamental modular structure to these system that SBA blocks will be used as fundamental components.

8.2.1 SBA Networks

As previously stated, the fundamental problem of SBA design is that single block algorithms are inherently tuned for a single class of data, e.g. pre-processing by image filtering is adjusted to the average expected noise levels. If the algorithm consists of a single block, it will have limited utility when presented with inputs that have dissimilar noise structures. One possible way to compensate for this effect is to use multiple SBA blocks embedded in a two-dimensional matrix, or network, structure. In this configuration, shown in Figure 8.3, a set of N SBA processes, P_n , are tuned to N classes of input

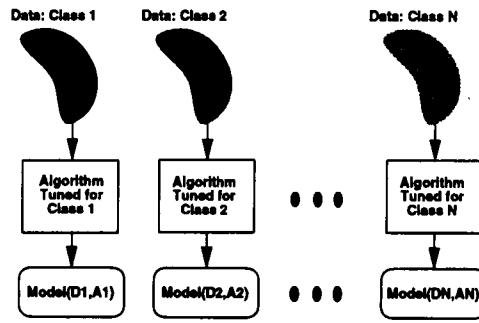


Figure 8.3: **SBA Network Algorithm:** A parallel model of algorithm design where an individual elements of a process matrix have been tuned for a particular class of input data.

data, D_k (Note: $M = N$). This type of system works well where the number of input data sets is limited and the characteristics of the input data are well-understood. It makes implicit use of parallelism and can be trivially mapped into a parallel architecture. It has no reconfiguration, or feedback systems.

This architecture can be simply extended by imposing a fully-connected matrix (input to process) on the system. In other words, each of the classes of input data, D_k , is routed to each of the processing components, P_n . This takes advantage of the full computational power of the two-dimensional network and will produce a model matrix with $N \times N$ elements. Shown in Figure 8.4, this model is highly parallelizable, but lacks feedback capability. Now that the characteristic of implicit parallelism has been imparted to the network, the paradigm must be extended to include reconfiguration and feedback capabilities. A functional feedback mechanism could be using a weighted scheme that is a function of the input data and the process operating on that data. An iterative mechanism that will adjust the weights based on the results of model evaluation completes the feedback and reconfiguration loop. In addition to these fundamental looping

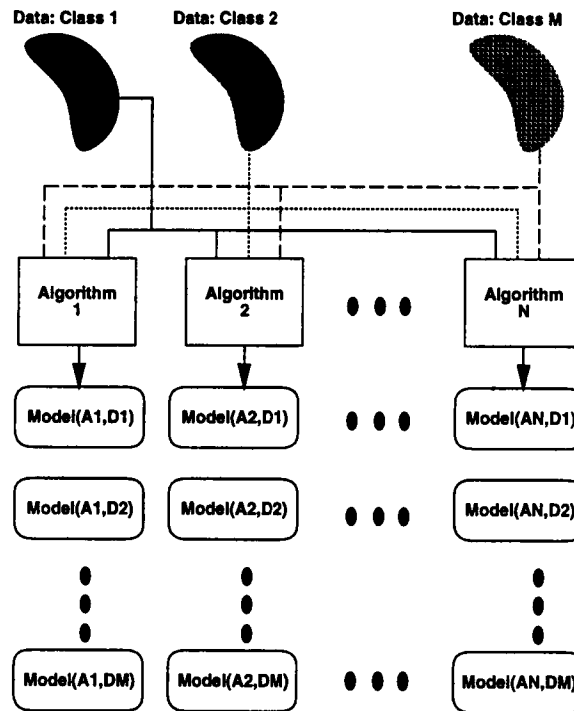


Figure 8.4: **SBA Fully Connected Network:** A more complex model of algorithm design where each data set is input to each process and the results area encoded into a model matrix.

constructs, it is useful to have a method of exploiting orthogonal information available when processing the data at different levels of resolution or from different channels of analysis in an optimal fashion. Figure 8.5 illustrates a system design model that has both parallization and dynamic reconfiguration capabilities. An operational theory for this model requires the development of several components. First, a mathematical model must be attached to this semantic model. This model must faithfully represent the weighting function, the evaluation functions, and the network reconfiguration capabilities that this design concept requires. The theory of Bayes [Pap84] is well suited to model the capabilities of this network.

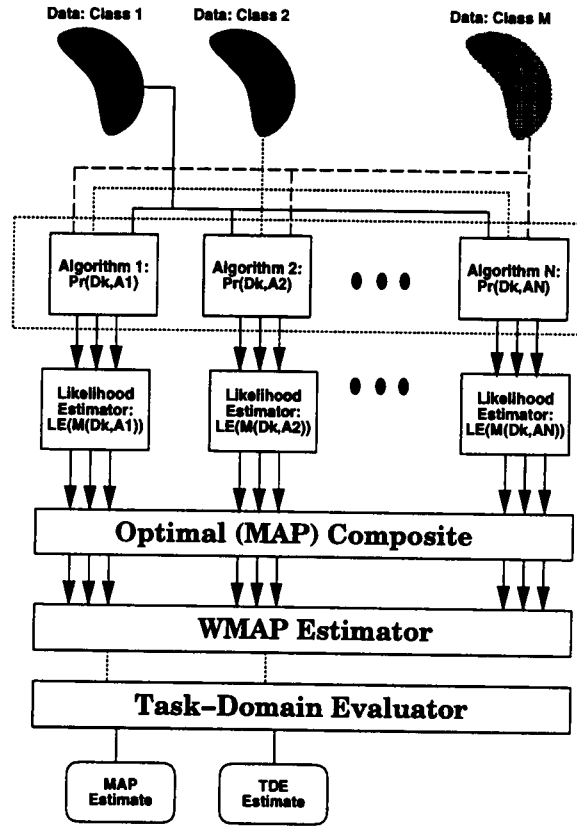


Figure 8.5: A **BAND Network**: A more complex model of algorithm design where each data set is input to each process and the results are encoded into a model matrix which is evaluated and used to adjust the optimal output.

8.3 Bayesian Algorithmic Network Design

The formal definition of Bayes' theorem and the fundamental building block of a BAND network is

$$Pr(M|D) = \frac{Pr(MD)}{Pr(D)}. \quad (8.3)$$

Stated verbally, this means that the probability of event M , given that condition D has occurred, is equivalent to the ratio of the probability of both M and D occurring to the

probability of event D . Usually this equation is presented as

$$Pr(M|D) = \frac{Pr(M)Pr(D|M)}{Pr(D)}. \quad (8.4)$$

Using this formulation, D is the input to the Bayesian system, i.e., an image. M is a set of parameters that describe some information deduced from the input, usually a representation of the desired output, i.e., the edges from an edge detector. The probability of D is called the Bayesian prior probability of the input. (Note: A common assumption is that $P(D) = \text{constant}$.) $P(M)$ is the prior probability of the parameter vector. And finally, $P(D|M)$ is the probability of the input given the model set, M . Given equation 8.4, a well documented technique to find the *maximum a posteriori*, or MAP estimate from a set of possible models M with respect to a given input, D is

$$M_{\text{map}|D} = \max_i [Pr(M_i|D)]. \quad (8.5)$$

Using standard optimization techniques [Par62,RRR83], distribution for the prior probabilities of the input models and a likelihood estimator for those models, the MAP solution is easily computed. Two fundamental problems exist in finding MAP estimates: (1) generation a probability density function that represents the Bayesian priors of the model, and (2) the construction of data-driven likelihood estimators that represent the principal part of Equation 8.5.

8.3.1 A Bayesian Model

Using the BAND model in Figure 8.5, the first part of the model is clearly a two-dimensional Bayesian probability network with structure,

$$Pr(M_{\text{map}}|D_k) = \frac{Pr(D_k|M(i, k)) Pr(M(i, k))}{Pr(D_k)}. \quad (8.6)$$

From equation 8.6, $M(i, k)$ is the model generated by process P_i when operating on input D_k . Again assuming the probability distribution of the input data sets is uniform, then the map estimate can be given by

$$M_{\text{map}|D_k} = \max_i [\ln \text{Pr}(D_k|M(i, k)) + \ln \text{Pr}(M(i, k))]. \quad (8.7)$$

Equation 8.7 and Figure 8.5 indicate, given an input data set of a particular class, a global MAP estimate is obtained from each of element of the available processing matrix provided an adequate likelihood estimator was provided. In addition, the user has full control over the two-dimensional Bayesian prior matrix, $\text{Pr}(M(i, k))$.

From here, one of the major problems of MAP estimation is explored. Even in cases where excellent likelihood estimation functions exist, there is no control on the quality of MAP estimation. In cases, where optimization is employed as a tool to obtain MAP estimates in a timely and efficient manner, the problem is further compounded by smoothness requirements of many optimization algorithms. Fundamentally, there is no guarantee that useful modeling information will be realized from the Bayesian estimator. In applications where a high level of confidence in the resultant estimations is demanded, task-domain knowledge provides additional constraints that will positively impact the results.

8.3.2 A Task-Domain Estimator

In this schema, models generated by each of the individual processes are model estimates, possibly MAP estimates in their own right, given input data and the process. Thus, a vector of M MAP estimates is available. A simple solution would be to do a global MAP estimation on the model probability matrix and set that to be the solution. However, this solution neglects two fundamental problems with MAP estimation:

- No effort is made to exploit orthogonal information that may be contained in a subset of the initial MAP estimates, and
- No explicit correlation between MAP estimation and correct results is implied.

These problems are addressed through the addition of task-domain constraints to buttress the integrity of MAP estimation and the generation of an optimal blending function that isolates complementary information from different channels of the networks and forces a more complete solution.

A Multi-Channel MAP Estimator

It is possible to consider each column of estimators in Figure 8.5 as a separate channel of multi-channel analyzer. Using this viewpoint, it is desirable to recombine this possibly orthogonal information into a form that represents more accurately the salient qualities of the input data rather than any characteristics imparted by the individual analyzers. Since the basis functions for internal representation were selected in the previous chapter (Representations for Active Stereo), the fundamental results and properties of that analysis are exploited here. In particular equation 7.9 which relates representations at differing scales, and equation 7.13 which is a simple scheme for combining a set of PEFR estimations into a linearly-weighted combination called an "N-blend".

Assume first that the per process results have been obtained. In this case, a set of N MAP Models for the input data set D_k have been derived. Associated with these N models is a vector of probabilities from the initial MAP estimation, see equation 8.7, that are used to generate a seed point for the coefficients vector of equation 7.13. Assuming

these probabilities are taken at equivalent scale, the coefficient vector seed is defined as

$$\alpha_i = \frac{\Pr(M(i_{\text{map}}, k))}{\sum_{i=1}^N \Pr(M(i_{\text{map}}, k))}. \quad (8.8)$$

The seed curve is generated by the application of the coefficient vector from equation 8.8 and the MAP models from the N applications of equation 8.7 to equation 7.13. Next, this seed curve, $C(\alpha, D_k)$ is input to another MAP estimator with uniform priors to return the optimal blended function. This function $C(\alpha_{\text{map}}, D_k)$ is obtained from

$$C(\alpha_{\text{map}}, D_k) = \max_{\epsilon} [\Pr(C(\alpha + \epsilon, D_k))]. \quad (8.9)$$

The tri-phasic function of the task domain estimator is to: (1) vary the parameters of the blended estimation so as to select an optimal blend from the initial MAP estimates; (2) Given the set of initial MAP estimates and the optimal blend, force a task-domain evaluation of each estimate; and (3) Select the optimal estimate from evaluation of the task domain cost functions. In step 3, it becomes desirable to depart from the theme of MAP estimations so that the verification of MAP estimates in a task-domain framework can help alleviate the afore mentioned problems of Bayesian estimators.

A Framework for Task-Domain Estimation and Verification

At this point in the process, the primary candidates for promotion are the N initial MAP estimates from the individual processes and the optimal blend obtained from equation 8.9. From here two classes of task-domain estimators are defined:

- *Inclusive*, and
- *Exclusive*.

Inclusive constraints can be modeled as continuous probabilities, $[0, 1]$ that represent some computational evidence that a particular model offers when compared to the input image. Exclusive constraints are designed to exclude models from consideration in cases where computational evidence supports these constraints; thus, an exclusive constraint is a highly nonlinear function, i.e. a step function, that takes the discrete values of 1 or 0. Further, exclusive constraints weight both the Bayesian priors and the likelihood estimators from equation 8.7, while inclusive constraints operate on the likelihood estimators.

Given a specification for X inclusive constraints and Y exclusive constraints and assuming independence among the probability relationships, The weighting model for inclusive constraints is

$$I^X(i, k) = \prod_{x=1}^X \Pr(I_x | M(i, k)), \quad (8.10)$$

and is

$$E^Y(i, k) = \prod_{y=1}^Y E_y(M(i, k)) \quad (8.11)$$

for exclusive constraints where the exclusion function is defined as

$$E_y(i, k) = \begin{cases} 1 & \text{if } E(M(i, k)) > 0 \\ 0 & \text{, otherwise} \end{cases} \quad (8.12)$$

Thus, the formal optimization function for a BAND network is the maximum of

$$M_{\text{wmap}|D_k} = \max_i E^Y(i, k) \left[\ln I^X(i, k) \Pr(D_k | M(i, k)) + \Pr(M(i, k)) \right], \quad (8.13)$$

the weighted MAP estimate of the initial model estimates; and,

$$C_{\text{wmap}|D_k} = \max_{\epsilon} E^Y(i, k) I^X(i, k) [\Pr(C(\alpha + \epsilon), D_k)], \quad (8.14)$$

the optimal blend which was seeded with equation 8.8. This BAND is specified formally as

$$M_{\text{band}} = \max \{M_{\text{wmap}|D_k}, C_{\text{wmap}|D_k}\}. \quad (8.15)$$

The BAND estimate from equation 8.15 has some interesting properties that MAP estimators lack. These properties are:

- The possibility of a null estimate.
- The inclusion of possibly orthogonal information from equation 8.14
- Inclusion of task specific information in the form of constraints into the estimation process.

8.3.3 A Simple Example

In this example, a simple multi-channel BAND estimator is constructed to demonstrate the benefits of this technique over standard MAP estimation. We begin with an initial estimate of a variable width line profile with unity area that is contaminated with uniformly distributed noise, shown in Figure 8.6. To illustrate the functional utility of the BAND network a simple neighborhood smoothing and edge detector combination are used to locate the point of square wave transition. (Identification of this point will specify the complete geometry of the square wave.) Clearly this scheme will work well in cases where signal-to-noise ratios are high; however, in low signal-to-noise profiles BAND estimation will significantly outperform traditional MAP estimation.

In our simple BAND network there are two process and Bayesian priors are assumed to be static and equal. Process A does not smooth, seeking maximal derivatives directly from the raw data. Process B implements a 5X1 neighborhood smoothing operation

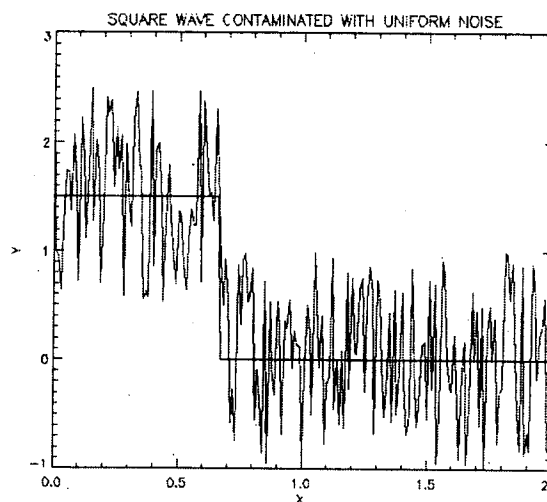


Figure 8.6: **Simple Square Wave:** shown with and without the application of uniformly distributed noise.

followed by the same edge detection procedure. A sample data set, such as Figure 8.6 is input to the network, and two transition point candidates, a and b , are output. From here, likelihood estimators are derived from the correlation of the model square waves specified by the candidate transition points, a and b , and the raw data. Once probability estimates are available, Equation 8.8 and Equation 8.9 are used to formulate an optimal blend for a and b . This operation produces another possible transition point, c .

Now, there are three estimates available. One from the unsmoothed detector, estimate a . One from the smoothed detector, estimate b , and one from the blending function, estimate c . The specification of task-level constraints is warranted. In particular, optimization will have difficulty with low signal-to-noise and will benefit from the added boundaries. In this example, the single constraint is an inclusive constraint (see equation 8.10) that weights the correlation-based likelihood estimate by a fraction of the perimeter. This constraint is added to control singular estimates of the square

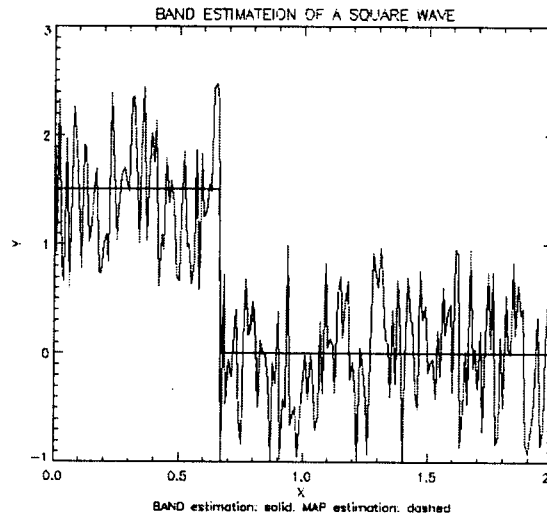


Figure 8.7: **Unit Area Square Wave:** contaminated with uniformly distributed noise in the range $[-1,1]$ with BAND and MAP estimates overlayed.

wave that can arise in highly noisy images. This perimeter constraint is

$$I^1(i, k) = \frac{x + \frac{1.0}{x}}{l}, \quad (8.16)$$

where l is a constant, 25.0 in these examples. Using equation 8.16 with equations 8.13 and 8.14, produces the weighted MAP estimations that are input into equation 8.15 to give a final result.

Figure 8.7 shows the results of application of this BAND network to a unity area square wave (amplitude 1.5) together with the results of the sample MAP estimation. Note from the figure, that both BAND and MAP estimations are correct for this high signal-to-noise profile. Figure 8.8 shows the same results with higher levels of uniformly distributed noise. The MAP estimates are wandering; however, the BAND network provides the correct result. Again, in Figure 8.9 the BAND estimation has proven superior to MAP estimation.

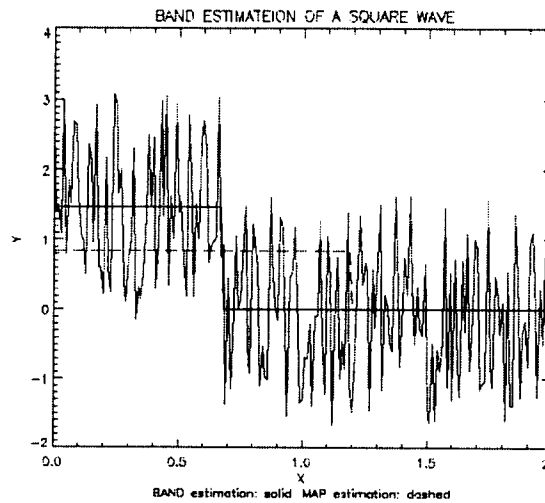


Figure 8.8: **Unit Area Square Wave:** contaminated with uniformly distributed noise in the range $[-1.6, 1.6]$ with BAND and MAP estimates overlaid.

8.4 Conclusions

A method for the construction of purposive vision algorithms has been proposed. This method, loosely based on Bayesian probability theory and Elliptical Fourier Representations, has several advantages over traditional MAP estimation. The integration of task domain constraints into the estimation processes provides an internal mechanism for relating simple physical relationships into an effective estimation and quality control procedure. Also, the existence of the null estimate highlights an important lack of traditional MAP estimation: the ability to reject posed estimates based on sufficient task-level knowledge. Finally, and most importantly, the ability to design estimators that include multiple layers of resolution or multiple channels of analysis in a parallel computational network offer superior performance with reduced design time.

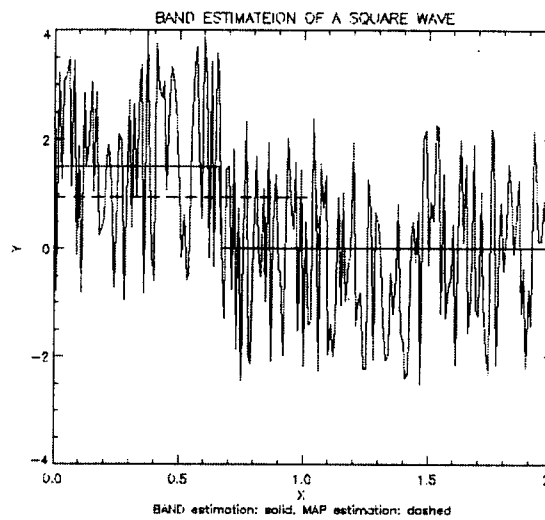


Figure 8.9: **Unit Area Square Wave:** contaminated with uniformly distributed noise in the range $[-2.5, 2.5]$ with BAND and MAP estimates overlayed.

Table 8.3: **BAND Square Wave Estimator (2 process algorithm):** Showing how a sequence of relatively simple algorithms can lead to very robust algorithms

SNR	0.75		0.90		0.45	
Trials	MAP	BAND	MAP	BAND	MAP	BAND
100	65	84	84	95	32	56
100	71	86	91	95	40	60
100	74	87	91	99	43	58
100	82	92	92	99	33	49
100	79	95	80	90	38	56
100	77	93	87	93	41	57
100	82	89	89	96	39	65

CHAPTER 9

Dual Arm Work Module

The Dual Arm Work Module (DAWM) is composed of two six degree-of-freedom hydraulic arms mounted on a five degree of freedom base. In reference to their mount points on the base, each arm has comfortable reach of about 80 inches directly ahead of the base mounting point. The only suitable location for mounting a stereo camera system was determined to be directly above the center of the base about 20 inches. This translated into a 135 in. offset from the stereo mounting point to the sensitive area of the robotic manipulator. In a single plane, the robot could move about 12.5 inches in any direction from that point without kinematic difficulty.

Grasping of small objects was the intended tasks. More specifically, objects between 1 and 4 inches were going to be cut with the "Jaws-of-Life" hydraulic cut tool. In addition, it was determined that it was necessary to cut objects along the entire 25 in. length that DAWM could transverse in the direction from base to tip. Putting these in terms of our stereo constraints with a rectangular work space gave the following:

- $l \approx 25.0$ inches.
- $w \approx 4.0$ inches
- $Z_d \approx 135.0$ inches.

Using PASS, it was found that

- $Z_{\min} \approx 98.0$ inches.

- $\Delta Z \approx 50.0$ inches
- $\Delta F \approx 8.0$ inches

Translating these variables into specific geometry gives:

- $b = 19.0$ inches.
- $\frac{f}{p} = 0.20$.

(Note: Actually focal-length for the system using our CCD cameras was calculated to be 48 mm.)

9.1 Implementation of a PASS System

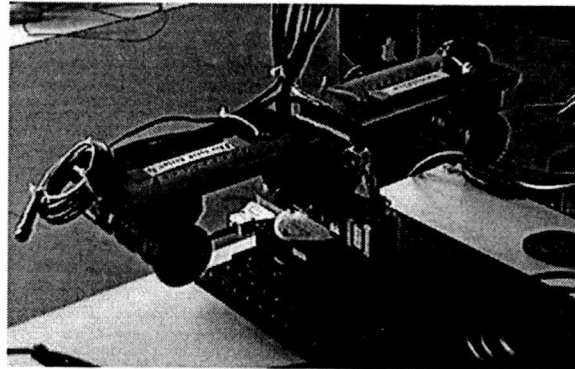


Figure 9.1: **Stereo Platform Design:** Stereo vision system designed using the PASS algorithm.

Stereo configurations generated with PASS are bounded by the parameters used in the synthesis(shown in Figure 9.1). In other words, if a design is optimized for a specific geometry, then objects must be present in the sensitive region before they will be imaged. In some more complex applications of this algorithm, a wider field-of-view will

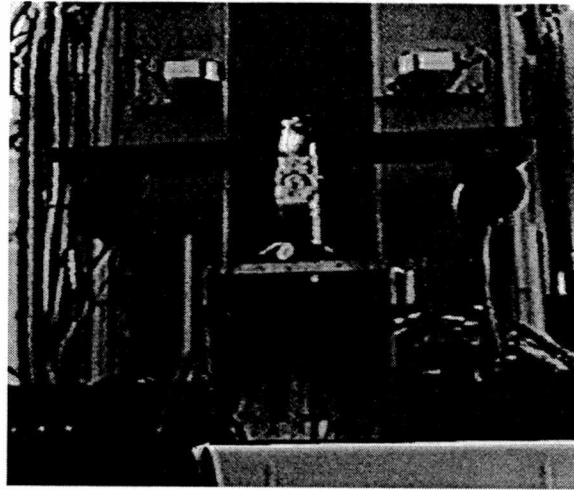


Figure 9.2: **Stereo Platform Design:** Stereo system mounted on the Dual Arm Work Module (DAWM) at the Oak Ridge National Laboratory.

need to be searched for the object while a PASS configuration is used for the ranging process. For the DAWM module, the constraints used were drawn directly from the task, cutting of small objects; however, the static geometry is not sufficient for the task of locating objects of interest from a large field-of-view. For this reason, a two degree of freedom rotational stage that allows pan and tilt motions was incorporated into the design. The unit was stepper motor driven with quoted accuracy of less than $\frac{1}{100}$ of a degree. It supported a full range pan with a kinematically restricted tilt of 45 degrees. The stereo unit is also shown mounted on the DAWM in Figure 9.2.

9.2 Simulations of the PASS Model

In order to provide some experimental validation of the model, a simulation of the stereo camera system was implemented. The model include spatial quantization typical of a standard CCD array, baseline uncertainty due to thermal expansion of the

mounting brackets and measurement error, disparity error caused by pixel localization problems in detection software, and uncertainty in focal-length caused by refocusing and aberrations in the optics [TT95b]. Parameter noise was assumed to be Gaussian and set to the following:

- Baseline Uncertainty: 0.0033 in. (std dev)
- Focal-length Uncertainty: 0.0025 in. (std dev)
- Pixel Uncertainty: 1 pixel (std dev)

This permits the assumption that maximum errors will occur at three standard deviations from mean [Sob94]; therefore, maximum errors were limited to 0.01 in. for baseline, 0.0075 in. for focal length, and 3 pixels for localization errors.

Figure 9.3 represents the maximum accumulated error for objects located 100 inches from the stereo array with baseline and focal-length perturbed. It shows that the maximum error distance was around 0.18in. (Note: That is total distance: representing all three principle directions) Figure 9.4 shows a similar plot with baseline and disparity perturbed. Maximum error was around 0.35 inches, well within the specified limits. Figures 9.5 and 9.6 respectively were contaminated with the same levels of noise; however the object points were located at a distance of 150 inches. Maximum combined distance error hovers at around 0.75 inches. While this figure is a little high, it does represent a 0.25 inch allowance in each of the principle directions. Also, it is important to acknowledge that most of the manipulation work will accrue in the 110-130 inch range. Also it is a measure of the maximum expected error, typical errors should have around half the magnitude.

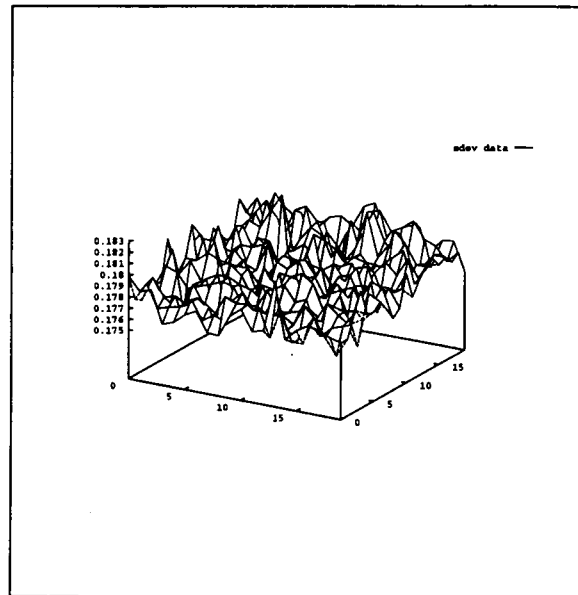


Figure 9.3: **Sensitivity Analysis:** Estimated range error for object distances of 100 in. with baseline and focal length perturbations.

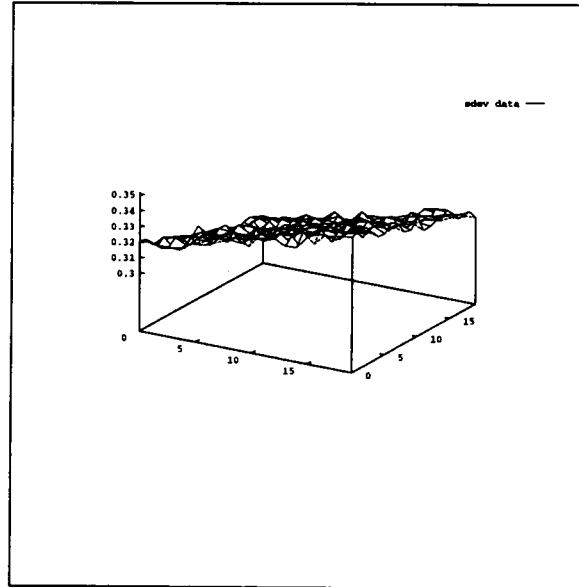


Figure 9.4: **Sensitivity Analysis:** Estimated range error for object distances of 100 in. with baseline and disparity perturbations.

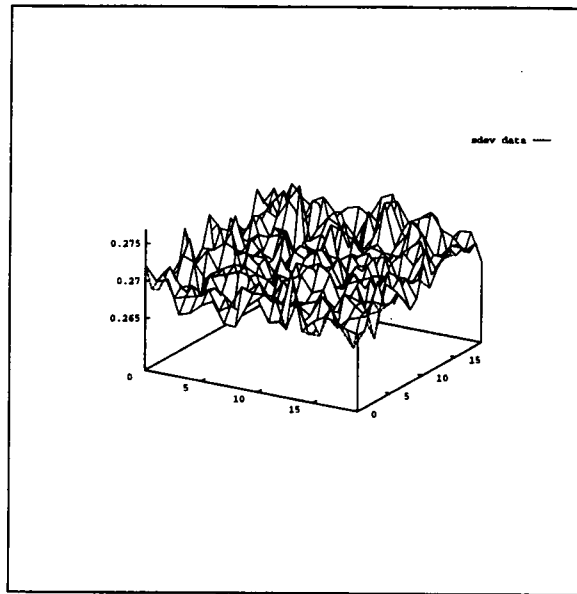


Figure 9.5: **Sensitivity Analysis:** Estimated range error for object distances of 150 in. with baseline and focal length perturbations.

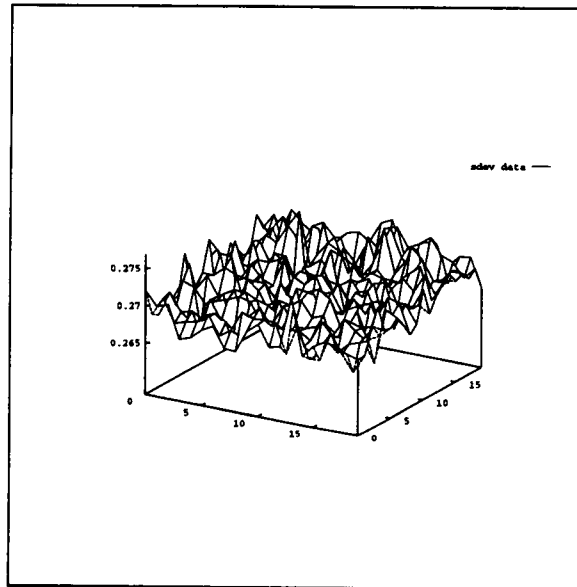


Figure 9.6: **Sensitivity Analysis:** Estimated range error for object distances of 150 in. with baseline and disparity perturbations.

9.3 Experimental Validation

Experimental data was obtained from the system by the placement of a planar target in view of the camera system. (Note: All stereo matching was done manually to minimize mismatching.) The target was randomly oriented and target spacing was not uniform. To test the accuracy of the system, three-dimensional range information was obtained from each target. Then, this data was used to find a best-fit in a least squares sense to a model of a plane. Residual error of the fit was obtained by using the derived model coefficients to compute the predicted values of one component of the model, and then subtracting the predicted from the actual.

The general equation of a plane in Cartesian coordinates is

$$ax + by + cz + d = 0; \quad (9.1)$$

however, this model needs to be rearranged to make it suitable for solution of the coefficients set in a least squares fit. This new model follows from a simple rearrangement of the terms in the form

$$z = -(mx + ny + p), \quad (9.2)$$

where $m = \frac{a}{c}$, $n = \frac{b}{c}$, and $p = \frac{d}{c}$. Given that we have N pairs of three-dimensional coordinates, these equations can be formulated in a matrix representation as

$$\mathbf{D}q = Z, \quad (9.3)$$

where $\mathbf{D}$ is the augmented, $N \times 3$ matrix of x and y values in equation 9.2; q is the 3×1 vector of coefficients of the planar model from the same equation, and Z is the remaining vector from the range triplet. Model coefficients are where found using the

normal equations [Str86,Pre88],

$$q = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{Z}. \quad (9.4)$$

Two sets of range data using the same three-dimensional reference point were obtained. Both sets were then fit to a plane using equation 9.4. (Note: approximately 50 range points were distributed over the entire stereo region.) Table 9.1 documents the statistics on the residual error. There, μ represents the mean absolute error instead of the average error, which in a least-squares approximation should be nearly zero. The standard deviation and the minimum and maximum values present in both data sets are also presented. A freeze frame from a video of the DAWM robot actuating an autonomous cut is shown in Figure 9.7.

Table 9.1: **Planar Model Coefficients:** Least-Squares coefficients for planar model of two sets of range data obtained from the same three-dimensional reference point.

μ	0.6563	0.5262
σ	0.8459	0.6699
min	0.0411	0.0024
max	2.0846	1.6863

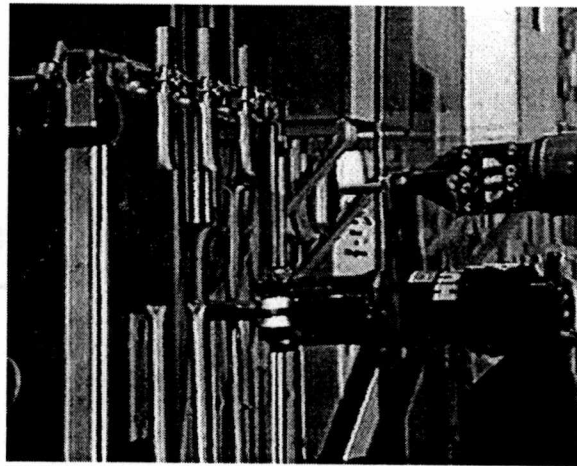


Figure 9.7: **Robotic Cutting Task:** Stereo vision system designed using the PASS algorithm is providing three-dimensional coordinates that direct the hydraulic cutting tool to its target.

CHAPTER 10

A Convergent-Axis Stereo System

In dynamic vision, static baseline systems with fixed lenses are not suitable for many tasks [KFS88, Kro89]. For instance, in the PASS stereo system designed in chapter 8, it is not possible to obtain optimal range estimates for a single point at some distance from the sensor arrangement; rather, one is forced to use the workspace definition that was included in the design phase to obtain estimates that are optimal for that entire workspace. Thus, range estimates at any single point in the workspace are typically sub-optimal. There exist a plethora of improvements over static baseline parallel axis systems [Cla92, OK93, AA88, Jai89, NP83]; however, in this chapter a dynamic-baseline, convergent axis stereo system platform is designed and built using the CASS design paradigm from Chapter 4. This system architecture will be built on the following principles:

- A weakly calibrated stereo system (where possible).
- Dynamic reconfiguration for optimal geometrical configurations.
- Automatic contrast and focus enhancement.
- Pipelined architecture for near frame rate execution.

10.1 Hardware Design

In contrast to the DAWM system built in Chapter 8, it was desired to build a desk-top stereo acquisition platform that was capable of convergence and made use of motorized lenses that were readily available. By definition the system was to operate on a single piece of optical breadboard measuring one foot wide by six feet long. The stereo platform would be buttressed against the base of the optical plate. A circle with a one foot radius, centered at four feet from the baseline coordinate system of the CASS platform, was set to be the desired workspace. This was necessary for operation of the automated camera features in their linear range of operation.

Using mainly surplus equipment from previous projects, i.e. pan/tilt units, linear motors, and variable zoom lenses, dictated some limits on the design. However, the available equipment was enough to implement a system with fundamental convergence and dynamic baseline capabilities. Figure 10.1 abstracts the kinematic capabilities of the system. The basic design consisted of mounting one of the motor units on a linear table and the other onto a base plate with the same vertical offset as the linear table. Next, the centers of the rotational stages were placed twelve inches apart on a piece of optical bread-board. With the linear table in the zero position, the baseline of the system was capable of motorized adjustment from ten inches to fourteen inches using the motor control software.

The optics of the system were manifested by two variable focus, zoom, and aperture lens systems. These systems were mounted on the rotational stages as shown in Figure 10.1. They replaced the fixed zoom lenses of the PASS system from Chapter 8. In a convergent stereo system, it is possible to optimize the zoom once the optimal

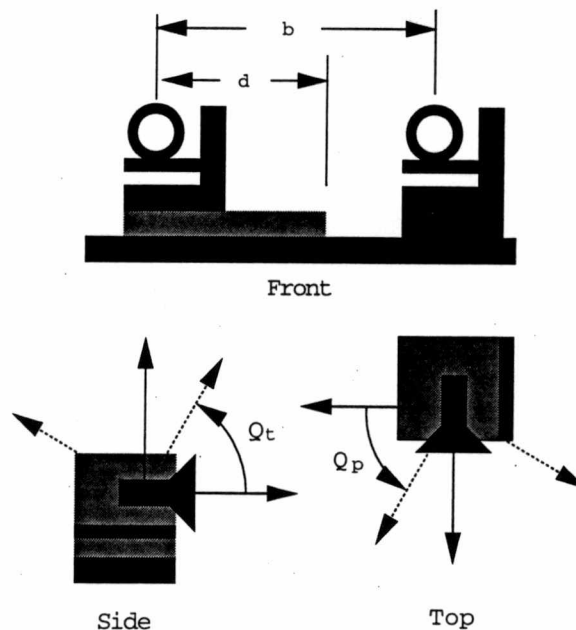


Figure 10.1: **Platform Concept:** Front, top, and side views of a conceptual platform with vergence and dynamic baseline capability.

convergence has been achieved. In this way, recovered range estimates are obtained with the maximum accuracy possible.

10.2 Computational Architecture

Computational muscle for the system was provided via a real-time image processing system controlled by a UNIX host. Each of the video feeds was piped into a Datacube image processor. The left camera made use of an MV200 system while the right camera used an older generation MV20. These systems are able to implement complex image processing functions at near frame rates. More complex, nonlinear or vector, processing was handled by transferring data from either of the MaxVideo (MV) boards to an IP860 vector processor with a fast bridge that bypasses the common VME backplane. Here,

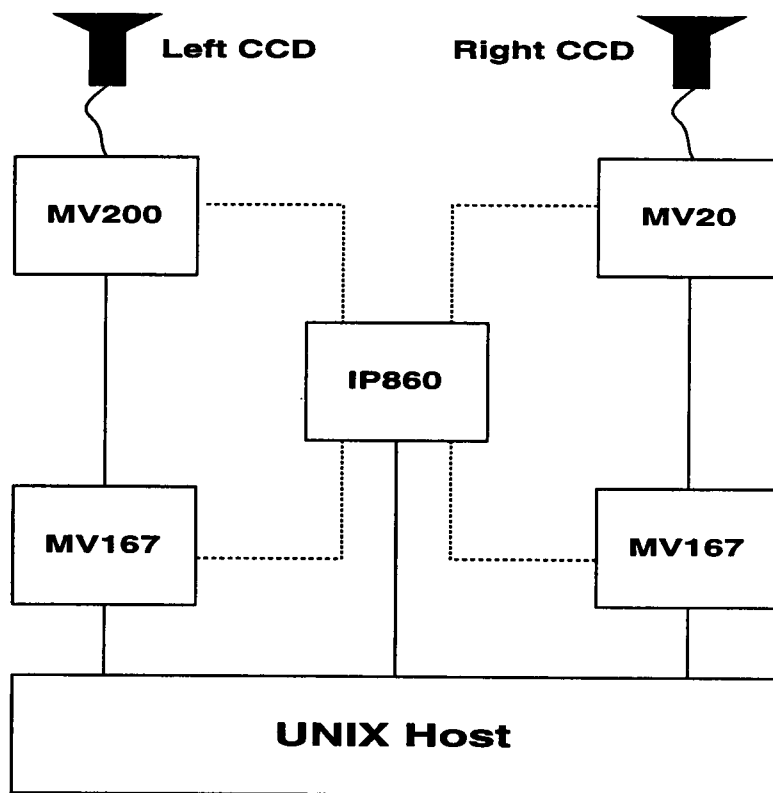


Figure 10.2: CASS Real time Computing Engine: A VME-based computing architecture for the implementation of low-level processing on convergent stereo systems.

tasks such as thresholding or integration can be implemented with high throughput. In cases, where the vector processor is occupied or motor control tasks need to be actuated, two Motorola MV167 (40MHZ) general purpose processor boards provide additional support. The main board is charged with managing the VME backplane, tasking the servo-lens motors, and communication with the UNIX host. The secondary processor serves as a general purpose processor. The complete computation system is depicted in Figure 10.2 Using this special hardware, three systems will be produced. First, a system for regional focusing will be implemented to improve local range estimates. Then a simple weakly calibrated, reflexive, auto-aperturing system will be derived and

tested for improving global image quality in highly unstructured lighting environments. Finally, the derivations from chapter 4 will be implemented into the motor and zoom lens control software to implement a fundamental capability of vergent stereo and auto-zooming behavior.

10.3 Pipelined Auto-Focus

A common manual operation in many vision systems is focusing. If objects are always placed at the same depth for viewing by a vision system, then a single focus setting is adequate. However, in a dynamic imaging platform, objects at many depths may be presented. In order to improve the usability of these designs, an auto-focusing functionality is critical. In addition, the ability to focus must be restricted to a subset, or region, of the input image. It does not make sense to focus an entire image, given a large depth of field in the scene; rather, using local regions, a gradient maximum technique was implemented to highlight regional features. Also, due to the dynamic nature of the uses of this imaging platform, it was desirable to formulate a set of requirements that will help to minimize the impact of the focusing algorithm on the user activity. These requirements are:

- Operates in < 5 cycles (6 Hz).
- Needs user input (the region of interest).
- Must settle quickly (no ringing).
- Weak calibration if possible.

10.3.1 An Optimal Technique

Tenenbaum [Ten71, TB76] has previously shown that a region of an image can be brought into crisp focus by maximizing

$$T(f) = \sum_R \left| \frac{\partial I(x, y)}{\partial x} \right| + \left| \frac{\partial I(x, y)}{\partial y} \right|, \quad (10.1)$$

where R is the region of interest from the input image $I(x, y)$. In active vision equation 10.1 is known as the Tenengrad function. The region of maximum focus can be found by

$$T'(f) = \max_f T \quad (10.2)$$

Thus, the best focus can be found by computing the gradients at a subset of the focusing positions and choosing the one that gives a maximal response to equation 10.2.

The lone input to this system is the region, R . Since auto-focus is considered to be an intermediate level behavior, with some input required from high-level control software, the origin of R is not a primary concern. This subsystem, given the input image stream and the region of interest, will simply return an optimal focus position to the high level control software.

10.3.2 A Real-Time Implementation

Auto-focusing behavior was relatively simple to implement given the presence of the real time hardware. Once triggered, by the loading of a region R into the the MaxVideo processors, the data pipelines would feed only the pixel information from the region R into the individual convolution units that performed the gradient enhancement. Then the local non-linear processor and the ALU of the Datacube hardware was used to perform the final components of the Tenengrad Function. This value, see equation 10.2

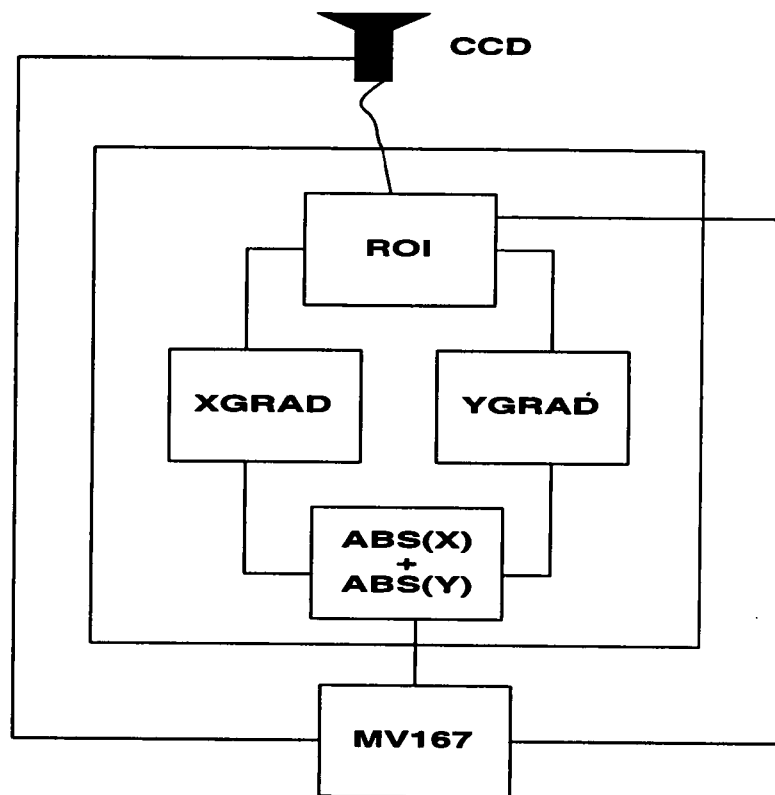


Figure 10.3: **CASS Auto-Focus Engine: Regional Focusing** via the Tenengrad function at (5 Hz).

was stored in the MV167 processor memory and the focus motor was advanced. A graphical version of the image pipeline is shown in figure 10.3

10.3.3 Results and Conclusions

A simple auto-focusing technique using constrained optimization was implemented on a real-time image processing system. Using windowed gradient maximization techniques, a region of interest was brought into crisp focus by searching for the focus position that gave the maximal response to equation 10.2. This subsystem relies on the ability of high-level control software to present a valid region of interest. It requires only an

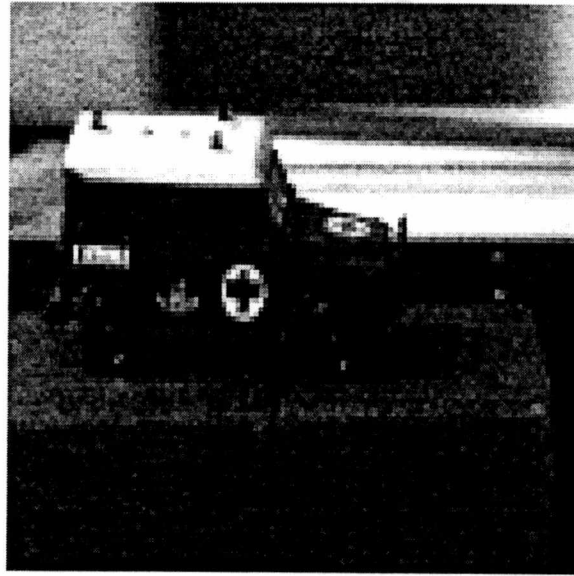


Figure 10.4: **Auto-Focused Image:** Toy truck with the region of interest centered on the cross.

operating range of parameters and does not require explicit calibration. Since the focus ring has much better settling times than the other lens systems it runs on the MaxVideo system at a rate of 2-5 Hz, depending on the system load. Figure 10.4 displays a sample object after auto-focusing was performed with the region of interest encompassing the red cross on the white circle of the toy ambulance.

10.4 Pipelined Auto Aperture

The textbook definition of illuminance is the “amount of light received per unit area on the image plane surface during exposure.” Since the aperture of a camera system is a ring of a certain diameter that restricts the amount of light impinging on the lens system, the illuminance is directly proportional to the square of the diameter. Therefore, the relative brightness of the scene can be controlled by changing the aperture settings.

A common problem in camera systems is the control of scene lighting to a level acceptable for automated processing. In tightly controlled lighting, it is possible to achieve good image contrast by fixing the aperture to a specific value; however, in dynamic environments this method of contrast setting can lead to many manual changes as lighting conditions vary throughout the day. A useful feature of a flexible desktop acquisition platform would be some sort of "real-time" automated control of the aperture to deliver an optimal image contrast. Since this level of behavior is typically reflexive in a biological system, it is desirable to make it function similarly in a machine system. This specification leads to certain requirements that must be followed in order for the system to be reflexive.

- Operates in 1-2 cycles (15 Hz).
- Runs without user knowledge (reflexive).
- Must settle quickly (no ringing).
- Weak calibration if possible.

10.4.1 An Optimal Technique

The first attempt at contrast enhancement was to try to mimic histogram equalization with the aperture. Technically, it is possible to change only the relative brightness of the image by changing the aperture setting; however, there does exist an aperture setting with maximal contrast. Thus if that setting could be found, the relative light levels would be controlled and the problem solved.

The problem is basically to seek the minimum difference between the current image histogram and the uniform histogram as shown in figure 10.5. The formal equation for

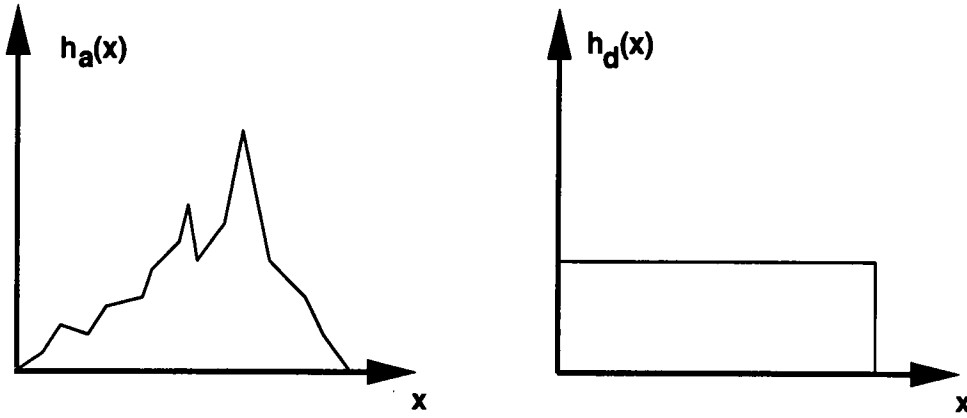


Figure 10.5: **CASS Auto-Aperture Concept:** Minimizing the difference between the current histogram and a uniform histogram given the input image.

the difference between the histograms is

$$e(x) = h_a(x) - h_d(x) = h_a(x) - C, \quad (10.3)$$

where $C = P/L$. P is the number of pixels in the image, and L is the number bins in the histogram. Squaring the error function produces

$$e^2(x) = h_a^2(x) - 2Ch_a(x) + C^2. \quad (10.4)$$

The minimal difference between the actual histogram, $h_a(x)$, and the desired histogram, $h_d(x)$, can be found by minimize the error term in equation 10.4 integrated across the entire histogram. Integration leads to the expression

$$\min_{\alpha} e^2(x) = \min_{\alpha} \sum_{x=0}^L [h_a^2(x) - 2Ch_a(x) + C^2], \quad (10.5)$$

where α is the aperture setting. This equation reduces to

$$\min_{\alpha} e^2(x) = \min_{\alpha} \sum_{x=0}^L h_a^2(x) \quad (10.6)$$

since the second and third terms are both constants. Notice that equation 10.6 states that the desired solution is the solution which offers the minimal energy in the histogram. While this solution is mathematically very attractive and can be solved by simple optimization techniques, two problems become immediately apparent. First the solution is iterative and relatively computationally intensive; it is not suitable for a reflexive system. Also, the large mechanical settling time of the aperture system, approximately two seconds, would send search times into tens of minutes. Either of these behaviors is highly undesirable, taken together a new approach is desperately needed.

10.4.2 A Weakly Calibrated Technique for Auto Aperture

Starting with the definition of illumination, as previously mentioned, an auto-aperturing system is built using information relative to the current device settings. Thus, given an aperture setting, a new setting will be computed that gives improved lighting characteristics in a single step. The computation is analytical and can be performed during the time allotted in the requirements.

Figure 10.6 depicts two different settings of aperture on the same camera system. It is known that the ratio of illuminance between the two settings is proportional to the ratio of the squared diameters, that is,

$$\frac{I_2}{I_1} = \frac{\pi r_2^2}{\pi r_1^2}. \quad (10.7)$$

A measure that is tightly correlated with image illuminance is the mean value of the histogram, μ . Thus, substituting for illuminance and reducing equation 10.7 gives

$$\mu_2 = \mu_1 \frac{r_2^2}{r_1^2}. \quad (10.8)$$

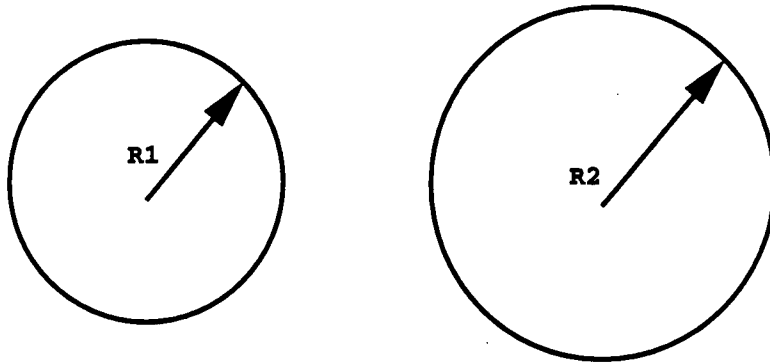


Figure 10.6: **Two Different Aperture Settings:** Illuminance is proportional to the square of the diameter.

Solving for r_2 produces

$$r_2 = r_1 \sqrt{\frac{\mu_2}{\mu_1}}. \quad (10.9)$$

The only remaining step is to decide what the correct value for μ_2 . This is the desired setting of the histogram mean, which is related to contrast.

$$r_2 = r_1 \sqrt{\frac{kL}{\mu_1}}. \quad (10.10)$$

It can be set, as in equation 10.10 to any value from the interval $[0, L]$. A good rule of thumb is to set it to $L/2$, e.g. $k = 0.5$.

10.4.3 Results

In general two conditions will be present in any image presented to the system. In the first kind the image intensity is too low. In Figure 10.7, the aperture setting on the input image was manually set to give a generally dark image. After the auto-aperturing algorithm was performed on the input image, the updated histogram displays a more distributed array of gray-level information. The next general class of input images shows an image histogram (solid lines) that is clearly saturated at the high intensity

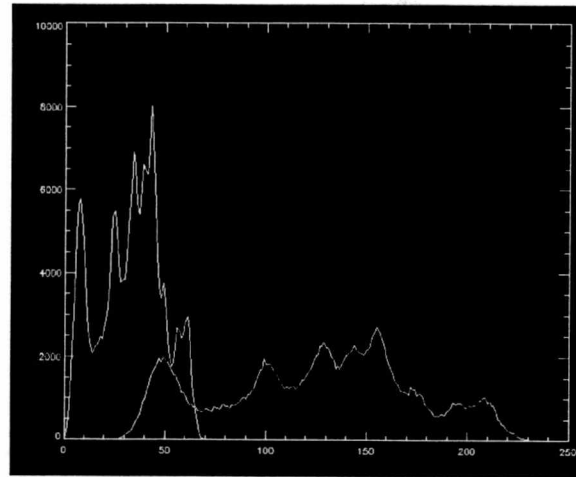


Figure 10.7: **Corrected Aperture Setting (Too Dark):** Aperture position is solved using a weakly calibrated model to give fast performance.

range. After auto-aperturing was run on the input system, the histogram has been redistributed to a more sensible structure. However, due to the approximation of the weak technique, the saturation problem cannot be completely undone. An example of type of arrangement is shown if Figure 10.8.

10.4.4 Conclusions

A weakly calibrated auto-aperturing technique was implemented and verified on a real-time processor system. The algorithm resided at at the level of reflexive behaviors and functions exclusively from the user if it is enabled. It is sub-optimal in the sense that it does not implement equation 10.6; rather, it relies on a model derived from the basic physical relationship of illuminance and aperture size. It makes no assumptions about the calibration of the aperture settings, which will have to be applied in most lens systems that do not support relative movements. Implemented on the MaxVideo processing boards, the auto-aperturing algorithm delivers corrected aperture settings at

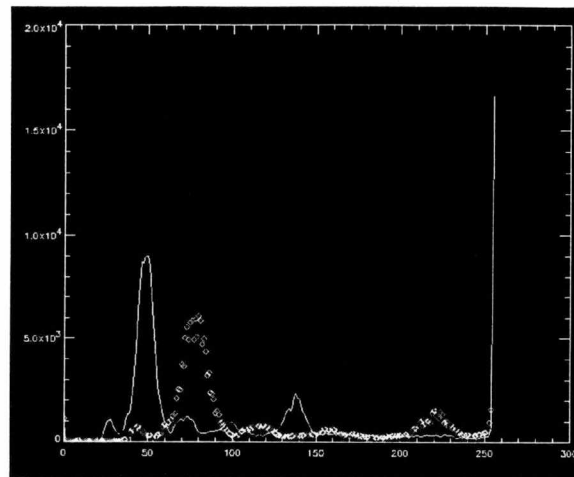


Figure 10.8: **Two Different Aperture Settings (Too Bright):** Aperture position is solved using a weakly calibrated model to give fast performance.

the rate of 10-15Hz.

10.5 Dynamic Regional Reconfiguration

The experimental verification of the mathematics derived in chapter 4 involved reconstructing the target grid shown in Figure 10.9. Essentially the 3-D coordinates of this plane are known to within the error tolerances of the optical plate. An estimate of each point location was computed using the following four step procedure:

- An optimal set of convergent stereo variables is computed; and the stereo platform is reconfigured.
- Auto-aperturing is performed.
- Auto-focusing is performed.
- A range point is computed.

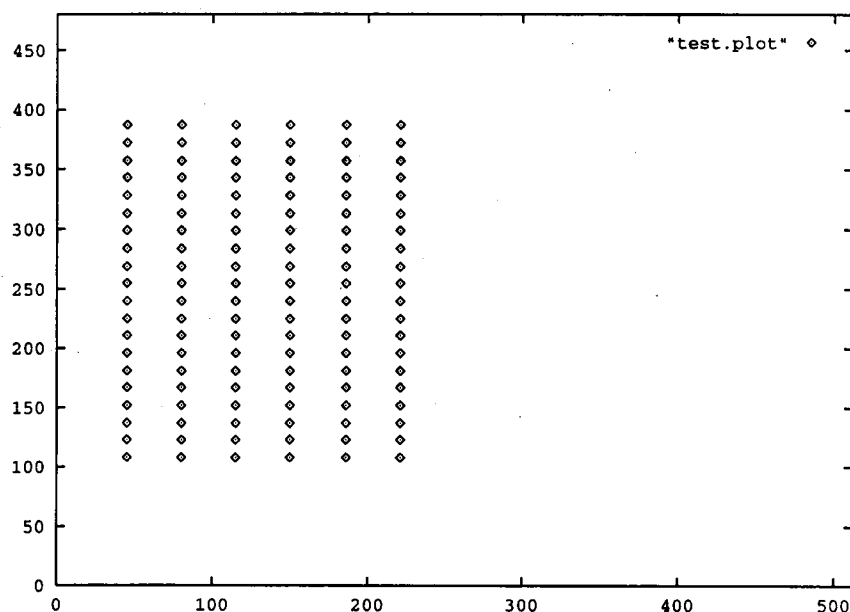


Figure 10.9: **Stereo Target Grid:** for the convergent stereo vision system.

Once estimates for all the target grids have been obtained. The residual error of a planar fit is computed and shown later in the document. Ideally, these results should demonstrate the superiority of convergent stereo systems over parallel-axis stereo systems; as well as experimental verification of the underlying mathematics.

10.6 Range Estimation

Using the same experimental setup from chapter 8, the accuracy of the convergent stereo system was analyzed. Again a randomly oriented target was placed in the field of view. Each scan point was manually targeted and the dynamic reconfiguration was

Table 10.1: **Planar Model Coefficients:** Least-Squares coefficients for planar model of two sets of range data obtained from the same three-dimensional reference point.

μ	0.317	0.482
σ	0.243	0.189
min	0.0011	0.0015
max	0.832	0.5321

used with a circular workspace region of two inches to improve speed performance. Once system reconfiguration had completed. Auto-focusing of each lens system was performed. Then, assuming perfect convergence, the values of the range estimates were computed. This data was then fitted to a plane using the same methods as chapter 9. Table 10.1 relates the statistics of the residual error. Again, μ is the mean absolute error since in a least squares fit the average error should be zeroed.

10.7 Conclusion

A CASS stereo platform, see Figure 10.10 has been implemented and tested. The platform was equipped with three abilities that give improved performance for dynamic stereo tasks over the PASS system of chapter 8. Auto focusing behavior provides the ranging system with crisp image sets for stereo matching. This ability is considered an intermediate level behavior since it requires the specification of a region of interest from high level control software. Generally this software ability is exercised after dynamic reconfiguration has been successfully completed and before range estimation has begun. Next, an auto-aperturing method was implemented using weak, or partial, information from the calibration model. This model has the tremendous advantage of taking only

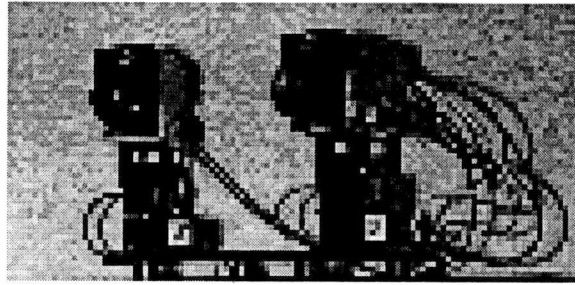


Figure 10.10: **CASS Platform Design:** Design sample without linear actuator to supply dynamic baseline capability.

two frame cycles to complete and offers a closed form solution to the aperturing problem. Finally, a dynamic reconfiguration scheme using the nonlinear optimization from chapter 4 was implemented on the control processors. When an objective region has been properly quantified, the dynamic reconfiguration process, handles the anti-symmetric convergence, the zooming characteristics and the baseline structuring in a quick optimization technique. Manual verification of range estimates proved to be superior to those of the PASS system in chapter 8 per unit baseline.

CHAPTER 11

Automated Synthesis of Structure from Cranial Imagery using BAND Networks

The process of attenuation correction in PET imaging is critical to quantitative studies of the human brain. Traditional computational methods rely on calculation of an attenuation map from the raw sinogram edge boundaries. These methods have the advantage of obviating the need for reconstruction that image-based attenuation corrections require as a precursor. However, in three-dimensional acquisitions or low signal-to-noise ratio, or low SNR, studies, sinogram thresholding methods may be problematic. While image-based attenuation correction may require additional reconstruction time, the use of prior information and structural relationships among neighboring images in a brain volume, may result in better detection and localization rates than in reduced SNR images, especially three-dimensional acquisitions. In this vein, an image-based brain attenuation correction algorithm has been implemented.

11.1 Introduction

This implementation consists of four major algorithmic components as shown in Figure 11.1. Volume segmentation is the first processing step. In this stage, the image volume is segmented into three separate regions, superior, inferior, and mid-brain images. The contour generation algorithm parameters change depending on the particular region of the brain to which an image belongs. For instance, contour generation starts in

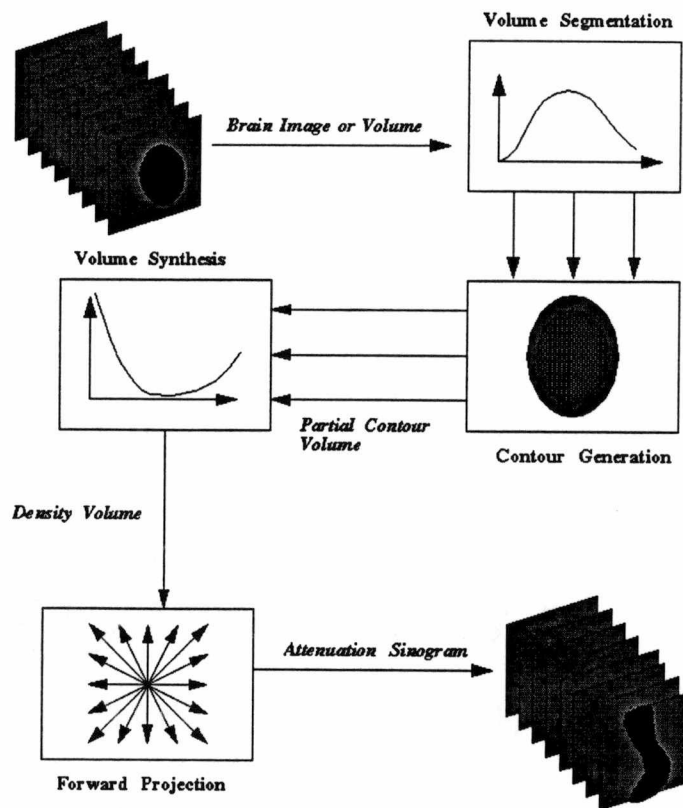


Figure 11.1: Block Diagram of Contouring Algorithm:

the mid brain images where the count rates are the highest and proceeds outward in both superior and inferior directions. In this way, initial guesses and parameter estimates are obtained from the images with the highest SNR. The second step uses a network of simple optimization procedures arranged in a competing hierarchical structure to locate the brain/skull boundary contour. Each optimization procedure is executed in order of increasing computational costs until an acceptable solution is located. If no acceptable solution is found, the image is discarded. Once the brain/skull boundary is located

and verified, the algorithm proceeds to locate the scalp/air boundary contour. When found, a skull layer is generated from the knowledge of the brain/skull boundary and the scalp/air boundary.

This step is repeated for each of the selected images from the input sequence. Upon termination of the contour generation algorithm, a specification of the soft tissue and bone structures in the brain has been synthesized for a proper subset of the input volume. Due to discarded images or the input parameters to the contour generation algorithm, the specification of a complete density volume, i. e. one density image for each emission image, is incomplete. The third stage of the algorithm, relies on interpolation as a means of generating the missing density images. Methods both for open-ended interpolation, e. g. an interval is bounded by only calculated density image, and normal interval interpolation exist. Again, these methods are dependent upon classification of the images from the initial volume segmentation in first step of the algorithm. The final step of the algorithm involves using the density parameters to generate a full volume of density images. These images consists of a solid generated from variable order elliptical Fourier representations. These solids represent a model of the brain consisting of a homogeneous soft tissue region together with an optional bone structure specification. The individual bone and soft tissue regions contain a constant value representing the attenuation coefficient of each of these structure. From here, these density images are forward projected and exponentiated to produce a complete attenuation map for the initial input volume. This attenuation map may then be input along with the original sinogram to an attenuation correction procedure.

11.2 Volume Segmentation

In a single head-first brain study, the image volume consists of an array of images that are arranged in order, starting with superior brain images and running through the mid brain to inferior brain images. While these initial and final images in the brain volume offer convenient starting points for the contour generation algorithm, they usually contain images that are the lowest in quality due to the reduced signal that is registered at these locations. What is desired, then, is to select as an initial point, images with the maximum count rates, or SNR. These images, usually located in mid-brain region, are segmented and labeled in the first step of the attenuation correction algorithm. In addition to this group of "high" SNR images, two other labelings are also produced. Superior brain images which are assumed to be bounding the mid brain region of the volume and linearly proceeding in the direction of the top of the skull possess the interesting quality that some planes for this segment of the brain volume may not contain a head signature at all; or, a few of the planes may possess such a weak signature that it is desirable to reject these images to prevent further processing. This is accomplished by direct ratio of the autocorrelation function from the edges of the image to the autocorrelation function from center of the total count mass (of the image). This measure gives some indication of the relative noise in the system.

Its value is thresholded and those images with deficient ratio are label as rejected, while the remaining images are passed for further processing. Due to the rapid bifurcation of structure in the inferior brain region, the only preprocessing done is the selection of an optimal image from which to begin processing of images from this region. This selection is a simple selection of the image with the maximum count rates. From

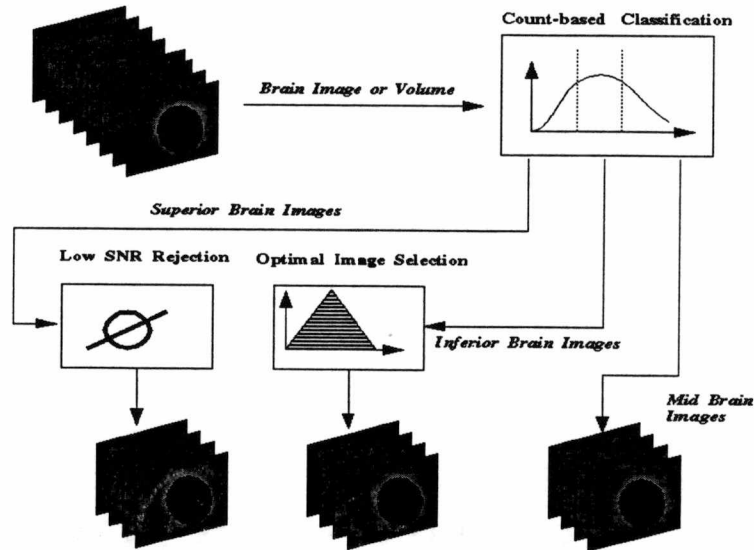


Figure 11.2: Volume Segmentation of PET Image Sequence:

here, these three labeled regions are passed in the order of (1) mid brain, (2) inferior brain, and (3) superior brain images, into the contour generation algorithm. Shown in Figure 11.2

11.3 Iterative Contour Generation Algorithm

Contour generation begins with a simple modeling of the tissue distribution of the brain as shown in Figure 11.3. As shown in the diagram, the tissue density from a simple transaxial cross section of the brain can be modeled as a single, solid region of soft tissue with a elliptical ring of bone embedded with it. Assuming this simple structure is an

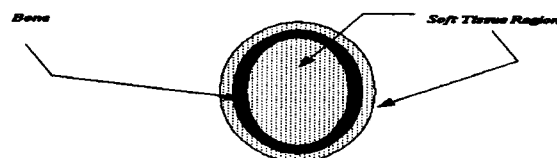


Figure 11.3: **Tissue Density Model for the Human Head:** The view is transaxial.

adequate model, three distinct structures are formed: (1) the uniform soft tissue region that is surrounded by the skull which represents the density of brain tissue (2) the elliptical ring which is representative of the skull density, (3) and the external uniform soft tissue region which is composed primarily of scalp tissue. (Note: The density of the background regions is by default assumed to be zero.) It is the task of the contour generation algorithms to produce a set of contours that can unambiguously reproduce this density image in a straightforward manner from the input set of images.

As mentioned, the input to the contour generation, at the slice level, is assumed to be an image of the distribution of biological markers within the brain. Independent of the nature of the marker, it can be assumed that the region of the skull will register substantially less activity than either of the remaining soft tissue regions. It is upon this assumption that the contouring algorithms are based. In them, two boundaries are sought: (1) the brain/skull boundary which is modeled as a step transition from

activity to non-activity radiating centrally about the center of mass of the entire image, and (2) the scalp/background boundary which is modeled in a similar manner after processing based on the results from step one. While these two contours have definite physical meaning, finding them is only an intermediate goal. Once localized, a synthetic map of attenuation values must be derived from them in a straightforward manner. As mentioned in Figure 6, the generation of a "mu" image with two distinct regions, one of bone and one of soft- tissue, is desired. This can be done is done from a simple algebraic combination of the two boundaries, together with the user specified values of skull thickness. In special cases which arise in practice, albeit infrequently, where one of the two possible contours was not localized or was rejected by the verification process are handled as follows: (1) when the scalp/background boundary is not present, the skull region is simple added as an direct extension of the skull/brain boundary with the appropriate mu value, and (2) when the skull/brain boundary is not present, the a uniform mu region is generated using the scalp/background contour as a boundary shell.

11.4 Boundary Localization in the Mid Brain:

Contour generation begins with the interval of images that represent the mid brain. For images from this region of the brain, the contour generation algorithm may proceed in two distinctly different manners. In first method, the existence of a seed point, or a starting boundary is assumed (Note: This is the normal mode of operation). This seed point, can be obtained from estimates of previous images, or estimates from algorithms that have previously run on the current image. Once an appropriate seed point has

been selected, the algorithms proceed by optimizing various energy functions until a suitable critical point can be located. After the optimizations have been performed, the generated boundaries are subjected to a series of simple evaluations to determine if the boundary is consistent with the images. If the boundaries are found to be consistent, they are passed to the next stage of processing. In the second method, no adequate seed points exist and the algorithm resorts to the generation of a seed point directly from the image data. Two distinctly different boundary generation algorithms exist. The first one attempts a direct segmentation of a discrete number of radial lines of response that intersect in the center of image mass. This segmentation is based on the locations of zero crossings in the second directional derivative parallel to these lines of response. The second algorithm implements an optimization of an energy function along these radial lines of response. Then energy function is a linearly decaying ratio of a windowed gradient and distance from image center. The optimal result the initial line of response is feed into the next optimization as the seed point. Once a complete set of boundary points have been generated in either method, a unique boundary curve can be synthesized from them directly.

Synthesis of both skull/brain and scalp/background boundaries in mid brain images proceeds in a direct manner. First, the skull/brain boundary is derived as mentioned in the previous paragraph. Once this has happened and the boundary has passed some simple verification steps (area, mass and center of mass checks), scaled versions of the boundary are used to generate simple filters for the image before processing to locate the scalp/background boundary is begun. This processing consists of the following steps: (1) The region of the image that is bounded by the curve is removed and replaced with a function of radially increasing intensity (This aids in the next boundary search). (2)

A simple filter is constructed from the original boundary that provides an upper limit on the search of the optimization for the scalp/background boundary. Values outside of this region are zeroed. Once this processing is completed, searching for the remaining contour proceeds in a manner similar to the first. Boundary Localization in the Inferior Brain

In the iterative contour generation algorithm, analysis begins with mid brain boundaries and is followed by inferior and superior boundaries, respectively. As previously mentioned, the inferior brain region is marked by a rapid bifurcation of brain structures as the brain stem becomes more prominent. For this reason, a traditional optimization as is done for mid brain searches for the brain/skull boundary is not possible. In order to overcome these inherent difficulties, the assumption that the skull/brain boundary does not deviate much as the brain images progress into the inferior brain region. Given this assumption is true, it allows the use of a single composite brain/skull boundary across the entire region, assuming corrections for minor translations and rotations from image to image. Therefore, using the last few contours generated from mid brain images as a road map, a single skull/brain contour is derived. Then, it is applied, with corrections, to each image in the inferior region. From here, the processing proceeds in the same manner as for mid brain images. The skull/background boundary does not experience the same erosion that the brain/skull boundary does; therefore, the same optimization routines are used here as in those from the middle of the brain.

11.5 Boundary Localization in the Superior Brain

Superior brain images are the last of the triple set of brain images to be processed. In this region of the brain, a steady, regular decay of both the SNR and the area of the both the brain and skull structures is experienced. For this region, processing of these images is further sub-divided into two separate regions. This is done for two reasons: (1) to gain an empirical measure of the rate of contraction in these brain structures, and (2) to insure that a pool of relevant images are used to derive brain contours for the superior brain structures. Again, segmentation of superior brain images into two regions is done by thresholding the total mass of the image intensities. Given this grouping into "Primary" and "Secondary" categories, processing is markedly different for each group. The "primary" interval, ideally, represents the images that are closes to the mid brain region, and it also should contain enough images so a suitable contraction parameter can be estimated. Processing in this region is different from processing of mid-brain images in only that a seed point is estimated for each "primary"-superior brain images. Once the set of "primary images has been analyzed, a rate of contraction, or scale factor, is calculated from the measured contraction of the area across this interval. This scaling factor is used to generate seed curves for images in the "secondary"-superior region.

Images from the "secondary" segment of the superior brain prove to be the most problematic for two reasons: (1) The SNR is rapidly decaying, as the area of the image is decreasing rapidly, and (2) some images from this interval may not have any usable signature from a true biological process at all. These images will be dominated by noise and should be rejected. Identification of the skull/brain boundary for images in this segment uses a algebraic contour solution generated from the "primary" interval scaled

with an appropriate multiple of the scale factor and corrected for minor translations and rotations. The scalp/background boundary is located in the same manner as that of the previous two regions except for two major points: (1) The optimization occurs on only the first few terms of the seed point curve and the resultant boundary curve is true ellipse regardless of the order of the seed point, and (2) The resultant boundary is cross-checked with the seed point to insure continuity of area, mass, and center of mass. If it fails these tests, the scalp/background boundary is rejected and a simple uniform soft tissue model is employed.

11.6 Partial Recovery Effects

Upon completion of the iterative contour generation algorithm, ideally a set of boundary curves for each of the input images should exist. However, in practice, many of the generated boundaries are discarded leaving a set of images that may have zero, one, or two boundary curves, as well as a set of images that have been rejected from processing all together due to low SNR. This is shown in Figure 11.4. In addition the first and last positions of the image volume may have intervals with relevant images, but no boundary curves for input into the attenuation correction algorithm. Additional processing to fill in these missing intervals is required.

11.7 Volume Synthesis

As previously mentioned, the final output of the contour generation algorithm is usually a partial set of contours for the brain images of the input volume. Thus, it becomes necessary to complete the boundary volume by "filling in" the missing bound-

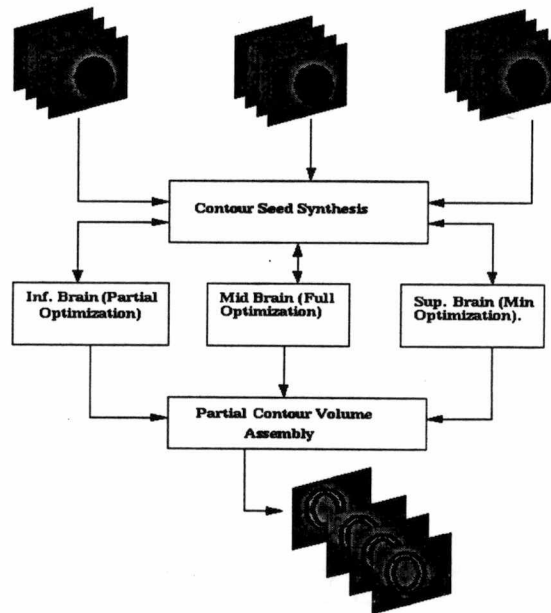


Figure 11.4: Iterative Contour Generation Algorithm:

ary sections. In estimating the missing boundaries across an interval, three distinct possibilities arise, as follows: (1) An interval with valid curves on both ends, or (2) an interval with a single valid curve in the superior brain, and (3) an interval with a single valid curve in the inferior brain region. (Note: the second and third cases are handled differently due to the disparity between the brain structures in the inferior and superior regions.) In Figure 11.5, the general process volume "interpolation" is outlined.

To begin with, an input volume is scanned for all the sequences of partial volumes that have valid curves on both ends. In these simple cases, a simple linear interpolation of the raw curve parameters is performed for each image in the interval that requires a curve estimate. This type of interpolation applies to mid, inferior, superior, and any of

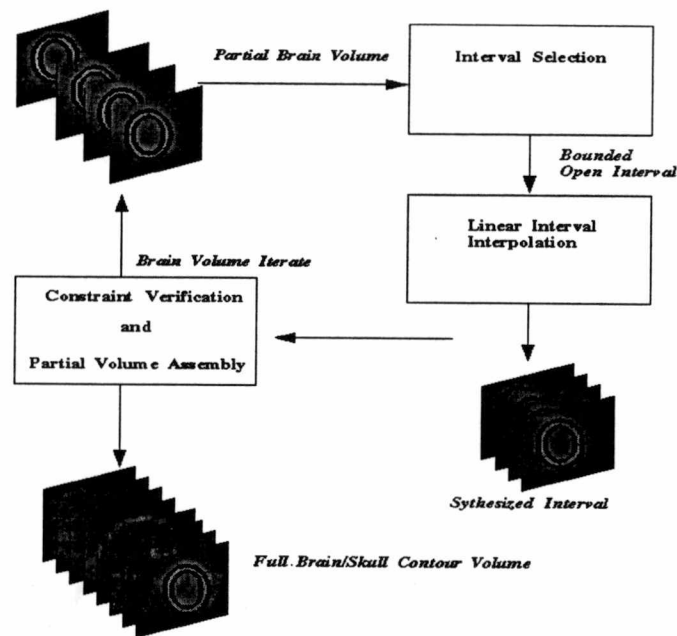


Figure 11.5: Volume Synthesis Algorithm:

their combinations in a uniform manner. The next possible type of interpolation is an open interval in inferior brain images. By definition this interval can be open only on the "downstream side" (as directions proceed from superior to inferior brain). If an interval of this type is isolated, the preceding sequence of images with valid curves is analyzed and an average representation is derived. This average representation is applied over the entire open interval with each individual curve being corrected for center of mass. The remaining possibility consists of an open interval in the superior brain. Again, by definition, this interval is open on the "upstream" side. For most images in this region, the geometry of the image is rapidly contracting as the images move up to the top of the brain. For this reason, the sequence of valid curves from the downstream region are

analyzed and from this sequence a line is fitted to the curve parameters. The slope of this line represents the contraction of the brain mass as images scan to the top of the brain, and the offset represents the translation of the center of mass. New curves for the open interval are calculated from the line parameters and corrected for actual center of mass.

Once all of the interpolations have completed, a fully generated volume of curves exists for the brain and the scalp (if the scalp boundary computed). However, this volume of curves must be converted into a density image for input into the forward projection process. Once this is done and the density images are forward projected, the attenuation map can be created.

11.8 Attenuation Sinogram Generation

Fundamentally, the process of constructing an attenuation sinogram from a two element curve model consists of four steps: (1) map each set of curves into a solid image containing the "mu" values for the particular tissue type, (2) perform a three-dimensional smoothing on the density image volume, (3) forward projection with reconstruction variables of (zoom, xoffset, yoffset, intrinsic tilt, and z rotation, and (4) exponentiation to produce a valid attenuation file. These steps are depicted in Figure 11.6.

The first step in producing an attenuation file is to take the set of curves produced by the contour generation code and making valid density images. Depending on the validity of the input curves, there are three distinct possibilities: (1) both scalp and brain boundary curves exists, (2) only a scalp boundary exists, (3) only a brain boundary exists. (Note: the fourth case where neither boundary exists is removed by the interpolation

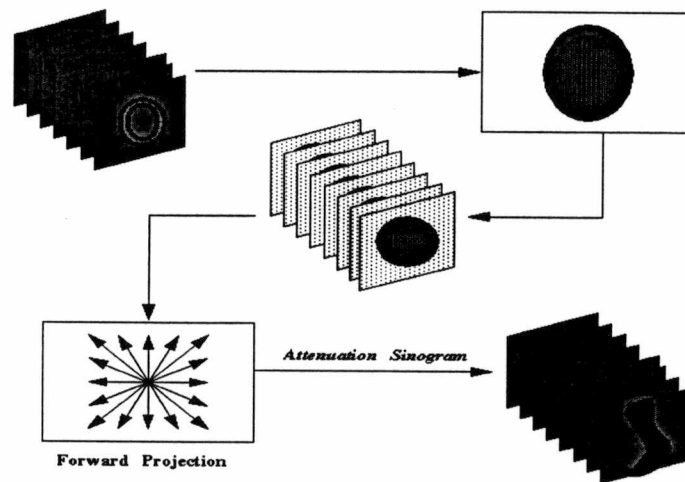


Figure 11.6: Attenuation Sinogram Generation:

step. In such cases where both scalp and skull boundaries exist, a uniform region of soft tissue constructed from the scalp curve is filled in with the density value for soft tissue. Then, a skull region of the specified thickness is overlaid onto the soft tissue region at the mid point between the scalp and brain curves. For the cases, where no scalp boundary is computed, a single skull layer of the specified thickness is grown onto the soft tissue region filled in from the brain boundary. And finally, when only a scalp curve exists, a skull layer of the specified thickness is overlaid onto the soft tissue region grown from the scalp contour, as in the case where both curves exist. Once this procedure has been applied to the entire volume, a three dimensional Gaussian with FWHM of 0.75 cm is convolved with the density volume

The final steps consist of a forward projection of the filtered density volume into a two-dimensional sinogram. This forward projection has the capabilities of transforming the forward projection sinogram into the space of the original sinogram by reversing the reconstruction variables mentioned previously. This results in a "log" sinogram of the density image, which is multiplied by the radial resolution and exponentiated before writing to disk. As a final note, all attenuation maps are written as two-dimensional volume mode sinograms.

11.9 Experimental Results

Figure 11.7 displays a reconstruction of a PET image with attenuation artifacts. Notice the tracer distributions along the transaxial centers exhibit a bowl shaped intensity distribution. Figure 11.8 displays the post-interpolation scalp/air and brain/skull boundaries from the iterative contour optimization procedure. Notice that the first four superior brain images, while possessing some signal, have been rejected by the algorithm due to low SNR. Figure 11.9 presents the contours after "filling" with the appropriate μ values. In this case a skull layer has been added in addition to the normal μ constant for soft tissue which constitutes the bulk of the brain density. The coronal and sagittal slices, in addition to the transaxial, are displayed in Figure 11.10. These planar views display the fidelity of the algorithm in reconstruction of both the skull layer and the recovery to the tissue density distribution of the human head from a system of tracer reconstructions. This algorithm is currently undergoing clinical trials to determine its utility in PET centers. Field trials of the system have been completed in Aarhus, Denmark, Lyon, France, and Detroit, Michigan.

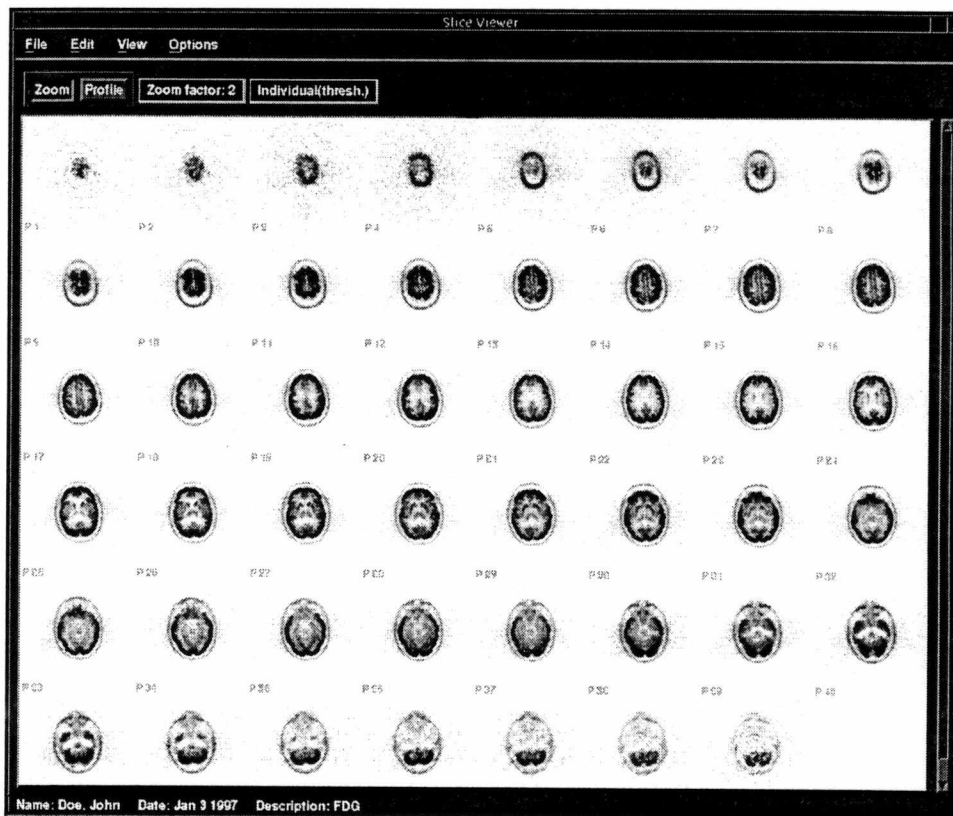


Figure 11.7: **Uncorrected Image Data:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.

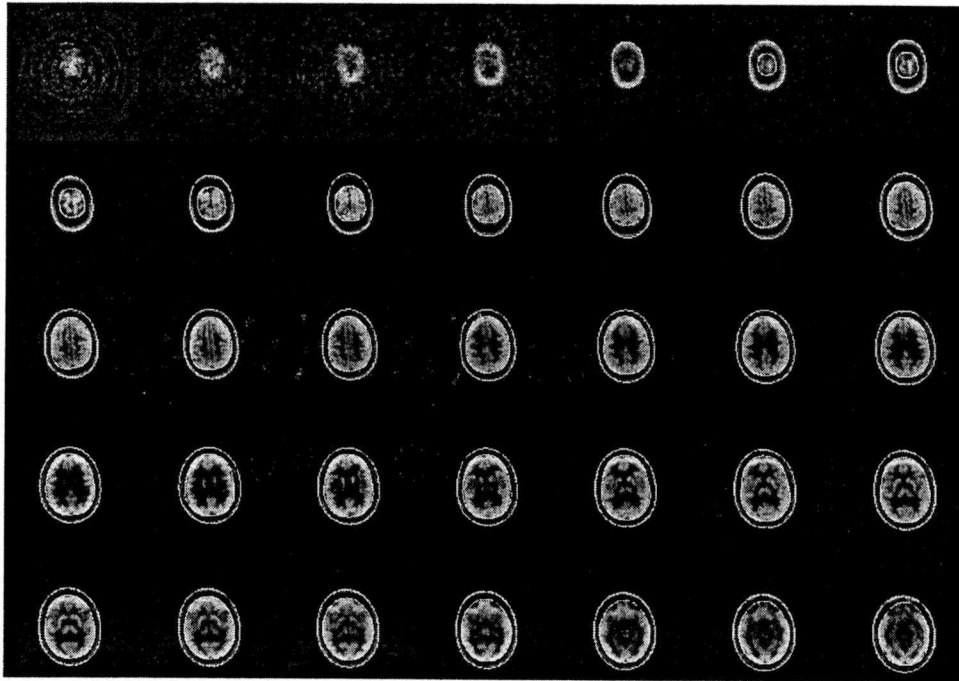


Figure 11.8: **Synthetic Contour Generation:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.

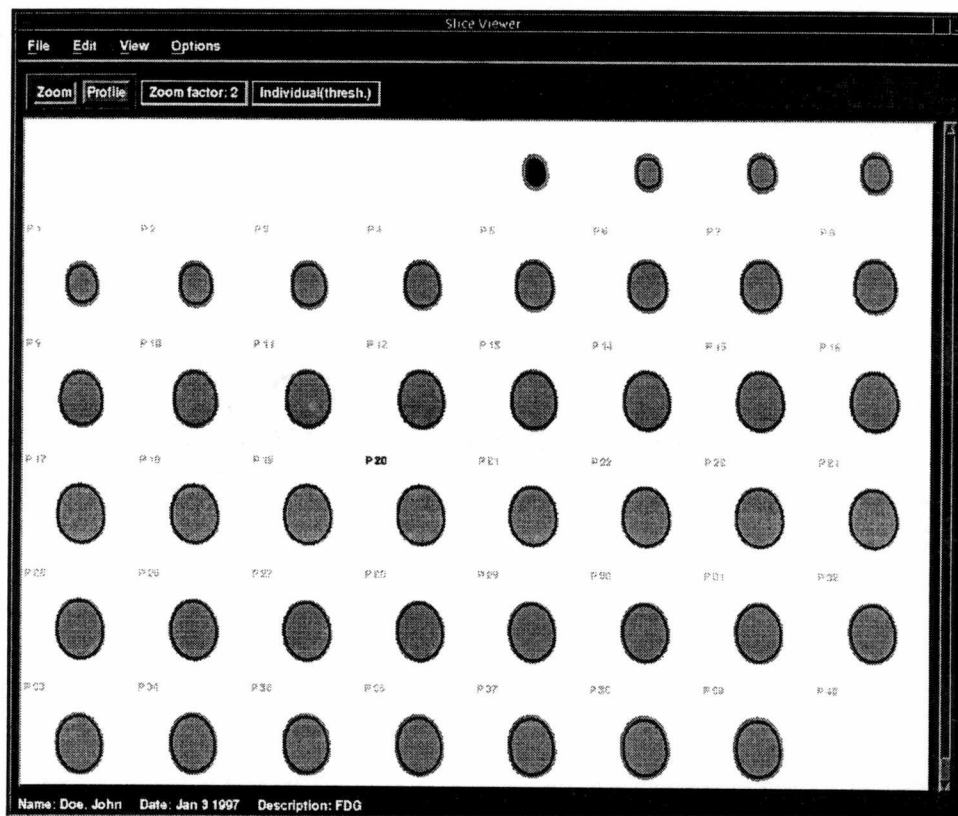


Figure 11.9: **Synthetic Tissue Density Map:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.

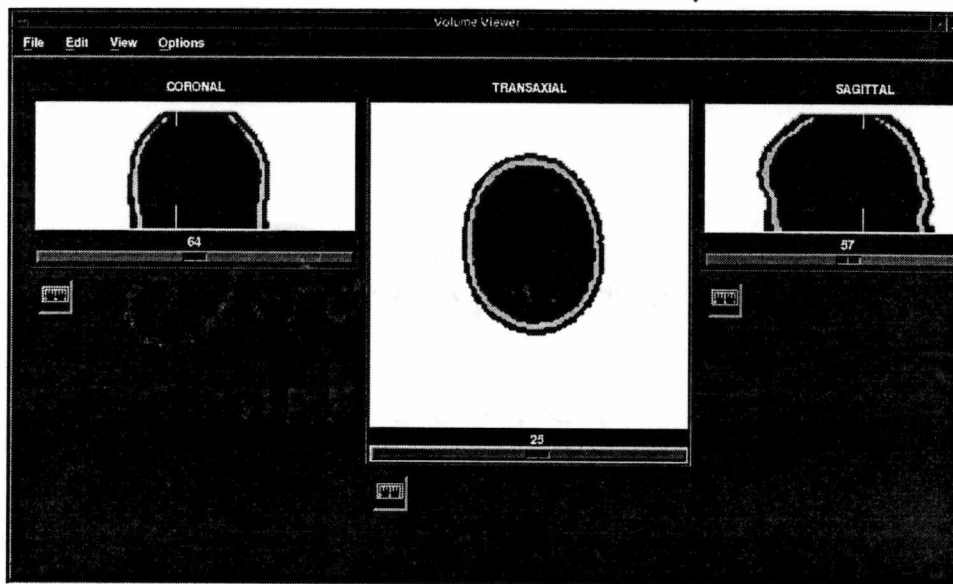


Figure 11.10: **Tissue Density Volume View:** A model of algorithm design where functional blocks are assembled sequentially to provide complex functionality in the assembled system.

CHAPTER 12

Discussion and Conclusions

The motivation for this work was to establish a set of procedures that could serve to initialize the design of a vision system in a task domain framework. The benefits of task-oriented design, from concept to completion, were manifest in reduced design complexity and a reliability increase in the context of a specific task. For applications that require accurate and robust estimates of input data in a timely fashion, superior design will lead to superior results. This *Task-Oriented Design Methodology*, or **TODEM**, offers a set of tools that facilitates purposive system design in a direct and undeniable manner. Under the TODEM architecture, four steps were used to implement a task-based design. The four steps were: (1) modeling the system (MODEL); (2) generating task requirements and integrating this knowledge into the system model (TRAIN), (3) formulation of cost functions using a goal list specification (FOCUS); and a final, possibly iterative evaluation of the integrated model to produce a task-based design (FIRE). With TODEM, three principal problems from general vision were studied in a purposive framework: (1) sensor geometry, (2) calibration, and (3) representation; as well as, the high-level algorithm design problem of how to implement solutions to these problems using task-domain, or purposive, constraints. The first of the problems that TODEM addressed was sensor design for a class of stereo vision platforms.

For the sake of brevity, the study of these platforms was limited to parallel-axis and convergent-axis stereo systems. An algorithm for *Convergent-Axis Stereo Synthesis*

(CASS) was outlined. In this algorithm two specific design problems were addressed, (1) static baseline convergent stereo, and (2) dynamic baseline convergent system. A triad of weapons was employed to solve this problem. A set of physical (bounding box) and task domain constraints along with the non-linear design equations was used with a nonlinear optimization procedure to arrive at an optimal design solution. A special case of the CASS algorithm, an algorithm for Parallel-Axis Stereo Synthesis (PASS), was presented. Armed with the design equations for this algorithm and knowledge of the work space, an optimal closed-form solution for the parallel stereo geometry was derived with minimal effort. The solution for each stereo configuration problem was offered via a minimum area technique that was proven to provide the maximum average accuracy over the entire workspace. As a byproduct of optimal design, the competing nature of accuracy versus workspace coverage was outlined and demonstrated using real test cases. Chapter 9 covers in detail a PASS system used in the localization and identification of small pipe networks at the Oak Ridge National Laboratory. While Chapter 10 describes a weakly calibrated CASS application for small object identification implemented at the University of Tennessee, IRIS Laboratory. Both of these stereo vision platforms were calibrated the second component of the TODEM architecture, a modified Two-Planes algorithm.

The heart of TODEM calibration is the well-known Two-Planes algorithm [IPS87]. It was modified to use task-domain information to provide data that fit the stability and accuracy requirements of a specific task in an optimal fashion. Since accurate calibration can be computationally expensive and numerically unstable, it should only be employed where absolutely necessary. Naturally, this type of selective specification fits directly into the TODEM framework. The calibration of two systems on opposite

ends of the localization/identification spectrum was explored using this philosophy. One system was designed to provide maximum accuracy over a limited field of view. This system demands a highly controlled level of physical and numerical stability. Using the PASS design algorithm and the calibration module, a stereo system that could recover depth with a maximum uncertainty of less than 3% as a function of distance with a standoff of 125 cm. was designed, prototyped, and tested. Accordingly, calibration and CASS were joined to design a system for small object identification which had no recovery requirements but had stringent real-time constraints. The former is discussed in Chapter 9 while the latter is demonstrated in Chapter 10.

Another component of TODEM design is adequate representation. This representation had three requirements that were derived from the fundamental capabilities of most active systems. *Parametric Elliptical Fourier Representations*, or **PEFRs** together with a formal system for implementing fundamental operators on them was presented. Using this representation has the innate advantage of designing model-based or template information directly into the initial sensor readings, thus task-specific foveation and attention schemas were encoded at the lowest-level of processing; as a consequence, object search times were drastically reduced. Chapter 10 records a CASS system for small object localization that was constructed using foveation and attention schemas for small object identification. Chapter 11 uses the same representational objects to perform object localization of cranial structures from Positron Emission Tomography (PET) imagery with demanding accuracy requirements.

Implementation of any subset of these three active vision design specifications will produce more robust visual estimators. However, for systems where robust operation across a wide range of input data and noise levels is required; or, where conflicting task-

domain requirements cannot be resolved, TODEM presents an architecture for visual algorithm integration using *Bayesian Algorithm Network Decisions*, or **BAND**. In this architecture, task-domain visual systems were designed to operate across a sequence of input data classes to provide unparalleled levels of robustness and accuracy of estimation. The key idea behind this plan is to use a series of related algorithms to exploit orthogonal information in a Bayesian framework to respond robustly to a wide range of input data. In this way conditional probabilities of how an specific visual algorithm responds to a specific class of visual input was encoded into the decision making process. In addition, a network history is maintained to provide reliability information derived from task-specific constraints at each network decision after training is completed. The reliability estimates were adjusted after each new data set is presented. Chapter 11 illustrates a 4x4 BAND model used in the accurate localization of structure in cranial PET images.

Three classes of visual systems where implemented and tested using the TODEM paradigm. These requirements of these problems were used to derive task specific (TRAIN) information that helps to reduce the design effort and improve reliability and robustness across a plethora of input data. A list of system goals was then used to formulate a task driven cost function (FOCUS) that is capable of expressing conflicting goals and inherent uncertainty relationships in a simple format. The system model, user specification and cost functions were then used in iterative evaluation to produce an optimal set of state variables for the system under test (FIRE). In each specific case, the effectiveness of the TODEM architecture was thoroughly tested and verified. It was also demonstrated that the range of support for TODEM applications is not limited to the class of vision systems using binocular stereo; rather, it can be applied to any spe-

cific vision problem that offers sufficient task-domain information. It does not address problems in general vision directly nor does it intend to. The implications of this theory in future work is considered in Appendix A.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [AA88] N. Ahuja and A. L. Abbot. Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for surface estimation. *IEEE Transactions on PAMI*, 15:1007-1029, 1988.
- [ABG89] N. Alvertos, D. Brzakovic, and R. C. Gonzalez. Camera geometries in image matching in 3-d machine vision. *IEEE Transactions on PAMI*, 11(9):897-915, 1989.
- [AC94] M. Adjouadi and F. Candocia. A stereo matching paradigm based on the walsh transformation. *IEEE Transactions on PAMI*, 16(12):1212-1218, 1994.
- [Alo88] J. Y. Aloimonos. Combining information in low-level vision. In *Image Understanding Workshop*, volume 2, pages 862-906, 1988.
- [Alo89] J. Y. Aloimonos. An extension of the marr paradigm. In *Image Understanding Workshop*, volume 1, pages 507-551, 1989.
- [Alo90a] J. Y. Aloimonos. Medusa synthesized. In *Image Understanding Workshop*, volume 1, pages 645-659, 1990.
- [Alo90b] J. Y. Aloimonos. Purposive and qualitative active vision. In *Image Understanding Workshop*, volume 1, pages 816-828, 1990.
- [Alo93a] J. Y. Aloimonos. Designing visual systems for purposive navigation. In Y. Aloimonos, editor, *Active Perception*, pages 47-102. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [Alo93b] Y. Aloimonos. Introduction: Active vision revisited. In Y. Aloimonos, editor, *Active Perception*, pages 1-18. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [Alo94] J. Y. Aloimonos. What i have learned. *CVGIP: Image Understanding*, 60:74-85, 1994.
- [BA94a] P. J. Burt and P. Anandan. Image stabilization by registration to a reference mosaic. In *Image Understanding Workshop*, volume 1, pages 425-434, 1994.
- [BA94b] P. J. Burt and P. Anandan. Real-time scene stabilization and mosaic contruction. In *Image Understanding Workshop*, volume 1, pages 457-465, 1994.
- [Baj88a] Ruzena Bajcsy. Active Perception. *Proceedings of the IEEE*, 76(8):996-1005, 1988.

- [Baj88b] Ruzena Bajcsy. Perception with Feedback. In *Image Understanding Workshop*, volume 1, pages 279–288, 1988.
- [Bal91] D. H. Ballard. Animate Vision. *Artificial Intelligence*, 48:57–86, 1991.
- [BB93] D. H. Ballard and C. M. Brown. Principles of animate vision. In Y. Aloimonos, editor, *Active Perception*, pages 245–282. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [Bes88] P. J. Besl. Active, Optical Range Imaging Sensors. *Machine Vision and Applications*, 1:127–152, 1988.
- [Bol89] R. M. Bolle. Data and model driven foveation. Research Report 67343, IBM, 1989.
- [Bro91] R. Brooks. Intelligence without reason. A. I. Memo 1293, MIT, 1991.
- [Bro94] C. M. Brown. Toward general vision. *CVGIP: Image Understanding*, 60:89–91, 1994.
- [BSS93] R. Bracho, J. F. Schlag, and A. C. Sanderson. Popeye: A gray-level vision system for robotics applications. Cmu-ri-tr-83-6, Carnegie Mellon University, May 1993.
- [Bur88] P. J. Burt. Smart sensing with a pyramid vision machine. *Proceedings of the IEEE*, 76:1006–1105, 1988.
- [Chr93] H. I. Christensen. A low-cost robot camera head. In H. I. Christensen, editor, *Active Robot Vision*, pages 69–87. World Scientific Publishing Co., Ltd., Hillsdale, New Jersey, 1993.
- [Cla92] J. J. Clark. Active photometric stereo. In *Proceedings of CVPR*, volume 1, pages 29–34, 1992.
- [CM94] H. I. Christensen and C. B. Madsen. A reply to “a computational and evolutionary perspective on the role of representation in vision”. *CVGIP: Image Understanding*, 60:103–108, 1994.
- [DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, New York, 1973.
- [DK93] F. W. DePiero and R. L. Kress. Camera Calibration Performed In Situ with Automated Analysis and Verification. In *Proceedings of the Fifth Topical Meeting on Robotics*, volume 1, pages 109–115, Knoxville, TN, 1993.
- [DNG92] F. W. DePiero, T. E. Noell, and T. F. Gee. Remote Driving with Reduced Bandwidth Communication. In *Proceedings of the Sixth Annual Space Operation, Application, and Research Symposium*, volume 1, pages 109–115, Houston, 1992.

- [Ede94] S. Edelman. Representation without reconstruction. *CVGIP: Image Understanding*, 60:95-97, 1994.
- [FC93] N. J. Ferrier and J. J. Clark. The hardvard binocular head. In H. I. Christensen, editor, *Active Robot Vision*, pages 9-31. World Scientific Publishing Co., Ltd., Hillsdale, New Jersey, 1993.
- [FF95] D. B. Fogel and L. J. Fogel. An introduction to evolutionary programming. In G. Goos, J. Harmanis, and J. van Leeuwen, editors, *Artificial Evolution*, pages 21-33. Springer Verlag, Berlin, 1995.
- [FGL87] K. S. Fu, R. C. Gonzalez, and C. S. G. Lee. *Robotic Control, Sensing, Vision, and Intelligence*. McGraw-Hill, 1987.
- [Fis94] M. A. Fischler. The modeling and representation of visual information. *CVGIP: Image Understanding*, 60:98-99, 1994.
- [Fre74] E. C. Freuder. A computer vision system for visual recognition using active knowledge. *MIT AI Lab Tech Rep*, (345), 1974.
- [Gib89] J. J. Gibson. *The Perception of the Visual World*. Houghton-Mifflin, Boston, 1989.
- [Goe67] J. W. Goethe. *Goethe's theory of colours*. Murray, London, 1967.
- [Goo68] J. W. Goodman. *Introduction to Fourier Optics*. McGraw-Hill, San Francisco, 1968.
- [Gra95] A. Grace. *Optimization Toolbox for use with MATLAB*. The Math Works, Inc., Natick, MA, 1995.
- [Gri93] W. E. L. Grimson. Why stereo vision is not always about 3d reconstruction. A. I. Memo 1435, MIT, 1993.
- [GW87] R. C. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley Publishing Company, Reading, MS, 1987.
- [HA92] J. Herve and Y. Aloimonos. Exploratory active vision. In *IEEE Conference on CVPR*, volume 1, pages 10-15, 1992.
- [HN91] R. J. Holt and A. N. Netravali. Camera calibration problems: Some new results. *Computer Vision, Graphics and Image Processing*, 54(3):368-383, 1991.
- [Hor89] B. K. P. Horn. *Robot Vision*. The MIT Press, Cambridge, MA, 1989.
- [IH90] K. Ikeuchi and M. Herbert. Task oriented vision. In *Image Understanding Workshop*, volume 1, pages 497-507, 1990.

- [IPS87] A. Izaguirre, P. Pu, and J. Summers. A New Development in Camera Calibration: Calibration of a Pair of Mobile Cameras. *International Journal of Robotics Research*, 6(3), Fall 1987.
- [Jai89] A. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [Jai94] R. Jain. Expansive vision. *CVGIP: Image Understanding*, 60:86-88, 1994.
- [Jar83] R. A. Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on PAMI*, 5(2):122-139, 1983.
- [JHTRT95] G. Ravichandran J. Horstmann T. Rasmussen and M. M. Trivedi. Robust target tracking through the integration of spatial and temporal filtering algorithms. submitted to *IEEE Transactions on PAMI*, May 1995.
- [KFS88] E. Krotkov, F. Fuma, and J. Summers. An agile stereo camera system for flexible image acquisition. *IEEE Journal of Robotics and Automation*, 4:108-113, 1988.
- [KG82] F. P. Kuhl and C. R. Giardina. Elliptic fourier features of a closed contour. *Computer Graphics and Image Processing*, 18:236-258, 1982.
- [KNI⁺85] Y. Kuno, H. Numagami, M. Ishikawa, H. Hoshino, and M. Kidode. Three-Dimensional Visual Techniques for an Advanced Robot System. *IEEE Journal of Robotics and Automation*, RA-3:11-16, March 1985.
- [Kro87] E. P. Krotkov. *Exploratory Visual Sensing from Determining Spatial Layout with an Agile Stereo Camera System*. PhD dissertation, University of Pennsylvania, Department of Computer Science, May 1987.
- [Kro89] E. P. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo*. Springer-Verlag, New York, 1989.
- [LH87] C. Lin and C. Hwang. New forms of shape invariants from elliptic fourier descriptors. *Pattern Recognition*, 20(5):535-545, 1987.
- [LL95] M. Li and J-M. Lavest. Some aspects of zoom-lens camera calibration. CVAP Technical Report 172, Kungl Tekniska Hogskolan, 1995.
- [Mar82] D. Marr. *Vision*. Freeman, San Fransico, CA, 1982.
- [MH80] D. Marr and E. Hildreth. Theory of Edge Detection. *Proceedings of the Royal Society of London*, A(207):187-217, 1980.
- [MJT93] E. Milios, M. Jenkin, and J. Tsotsos. Design and perfomance of trish, a binocular robot head with torsional eye movements. In H. I. Christensen, editor, *Active Robot Vision*, pages 51-68. World Scientific Publishing Co., Ltd., Hillsdale, New Jersey, 1993.

- [Mor81] H. P. Moravec. *Robot Rover Visual Navigation*. UMI Research Press, Ann Arbor, Michigan, 1981.
- [MT94a] S. B. Marapane and M. M. Trivedi. Binocular robotic head for active vision research. *OE Reports*, 12(120):1,8, December 1994.
- [MT94b] S. B. Marapane and M. M. Trivedi. Robot convoying. *OE Reports*, 12(120):1,8, December 1994.
- [MT95] S. B. Marapane and M. M. Trivedi. Multi-primitive hierarchical (mph) stereo system. *IEEE transactions on PAMI*, 16:227-240, March 1995.
- [MTL93a] S. B. Marapane, M. M. Trivedi, and N. Lassiter. A Novel Graphical Simulation and Animation Environment for Integrated Sensory-Motor Systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS) Conference*, July 1993.
- [MTL93b] S. B. Marapane, M. M. Trivedi, and N. Lassiter. A Real and Virtual Robot Head for Active Vision Research. In *Proceedings of the Sensor Fusion V Conference*, November 1993.
- [ND88] B. Noble and J. W. Daniel. *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [Nom92] Y. Nomura. Simple Calibration Algorithm for High-Distortions-Lens Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, November 1992.
- [NP83] H. K. Nishihara and T. Poggio. Stereo Vision for Robotics. In *First Int. Symp. Robotics Res*, pages 489-505. IEEE Press, 1983.
- [OK93] M. Okutomi and T. Kanade. A Multiple-Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:353-363, 1993.
- [OKK73] S. H. Ohtheru, H. Kobayashi, and T. Kato. Eyes of the wabot. In J. T. Tou, editor, *Learning Systems and Intelligent Robots*. Plenum Press, New York, 1973.
- [OW83] A. V. Oppenheim and A. S. Willsky. *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1983.
- [Pap84] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc, New York, 1984.
- [Par62] L. A. Pars. *An Introduction to the Calculus of Variations*. John Wiley & Sons Inc, New York, 1962.

- [PE93] K. Pahlavan and J. O. Eklund. Heads, eyes and head-eye systems. In H. I. Christensen, editor, *Active Robot Vision*. World Scientific Publishing Co., Ltd., Hillsdale, New Jersey, 1993.
- [Pla92] P. J. Plauser. *The Standard C Library*. Prentice-Hall, Englewood Cliffs, NJ, 1992.
- [PP93] J. R. G. Pretlove and G. A. Parker. The surrey attentive robot vision system. In H. I. Christensen, editor, *Active Robot Vision*, pages 89–107. World Scientific Publishing Co., Ltd., Hillsdale, New Jersey, 1993.
- [Pre88] W. H. Press. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1988.
- [PUE93] K. Pahlavan, T. Uhlin, and J-O Eklund. Active vision as methodology. In Y. Aloimonos, editor, *Active Perception*, pages 19–46. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [RB90] R. D. Rimey and C. M. Brown. Selective attention as sequential behavior: Modeling eye movements with an augmented hidden markov model. In *Image Understanding Workshop*, volume 1, pages 840–849, 1990.
- [Ric96] S. J. Rich. Autonomous Navigation of a Mobile Platform Using an Inverse Perspective Stereo-Based Vision System and Auxililary Sensors. Master's thesis, University of Tennessee, Knoxville, 1996.
- [Rob91] D. Robinson. The mechanics of human smooth pursuit eye movements. *J. Physio.*, 180:559–591, 1991.
- [RRR83] G.V. Reklaitis, A. Ravindran, and K. M. Ragsdell. *Engineering Optimization: Methods and Applications*. John Wiley & Sons Inc, New York, 1983.
- [SD92] L. H. Staib and J. S. Duncan. Parametrically deformable models. *IEEE Transactions on PAMI*, 14(11):1061–1076, 1992.
- [SG94] G. Sandini and E. Grosso. Why purposive vision. *CVGIP: Image Understanding*, 60:109–112, 1994.
- [SGGT93] G. Sandini, F. Gandolfo, E. Grosso, and M. Tistarelli. Vision during action. In Y. Aloimonos, editor, *Active Perception*, pages 151–190. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [Sob94] I. M. Sobol. *A Primer for the Monte Carlo Method*. CRC Press, 1994.
- [SS90] V. Salari and I. K. Sethi. Feature point correspondence in the presence of occlusion. *IEEE Transactions on PAMI*, 12(1):87–91, 1990.
- [SS93] M. J. Swain and M. A. Stricker. Promising Directions in Active Vision. *International Journal of Computer Vision*, 11:109–126, 1993.

- [Sta70] P. A. Stark. *Introduction to Numerical Methods*. Macmillan Publishing Co., Inc, New York, 1970.
- [Str80] G. Strang. *Linear Algebra and its Applications*. Academic Press, New York, NY, 1980.
- [Str86] G. Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, Wellesley, MA, 1986.
- [TB76] J. M. Tenenbaum and H. G. Barrow. Experiments in interpretation-guided segmentation. *Stanford Research Institute Tech. Note*, (123), 1976.
- [TB94a] M. J. Tarr and M. J. Black. A computational and evolutionary perspective on the role of representation in vision. *CVGIP: Image Understanding*, 60:65-73, 1994.
- [TB94b] M. J. Tarr and M. J. Black. Response to replies. *CVGIP: Image Understanding*, 60:113-118, 1994.
- [Ten71] J. M. Tenenbaum. A laboratory for hand-eye research. In *IFIPS*, pages 206-210, 1971.
- [TG74] J. T. Tou and R. C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley Publishing Company, Reading, MS, 1974.
- [Tha93] S. M. Thayer. On-Line Stereo Vision and Graphical Interface for Decontamination and Decommissioning Applications Using the Advanced Servo Manipulator. In *Proceedings of the Fifth Topical Meeting on Robotics*, volume 1, pages 287-294, Knoxville, TN, 1993.
- [Tol62] G. P. Tolstov. *Fourier Series*. Pentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962.
- [Tsa89] R. Y. Tsai. Synopsis of recent progress on camera calibration for 3D machine vision. In O. Khatib, J.J. Craig, and T. Lozano-Perez, editors, *The Robotics Review*, pages 147-159. MIT Press, 1989.
- [Tso90] J. K. Tsotsos. A complexity level analysis of vision. *Behavior and Brain Science*, 13:423-455, 1990.
- [Tso94] J. K. Tsotsos. Representation without reconstruction. *CVGIP: Image Understanding*, 60:95-97, 1994.
- [TT95a] S. M. Thayer and M. M. Trivedi. Designing robot heads using task-domain constraints. In submitted to *IEEE Conference on CVPR*, 1995.
- [TT95b] S. M. Thayer and M. M. Trivedi. Residual uncertainty in three-dimensional reconstruction using two-planes calibration and stereo methods. *Pattern Recognition*, 28(7):1073-1083, 1995.

- [WCH92] J. Weng, P. Cohen, and M. Herniou. Camera Calibration with Distortion Models and Accuracy Evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:1095-1099, November 1992.
- [WF92] P. Whaite and F. P. Ferrie. Uncertain views. In *IEEE Conference on CVPR*, volume 1, pages 3-9, 1992.
- [YS63] L. Young and L. Stark. Variable feedback experiments testing a sampled data model for eye tracking movements. *IRE Trans. Human Factors in Enginerring*, 4:218-227, 1963.

APPENDICES

APPENDIX A

Future Research Possibilities

Active system is a complex and rapidly changing field in the infancy of its development. This work has addressed a limited class of problems associated with stereo vision; even while hinting at the possibility that the TODEM architecture can be applied to a broader range of active vision problems, it has not developed a broad-based theory of active visual system design. And it is in this vein that many research opportunities and extensions to the TODEM theory exist.

A.1 A General Framework for Digital Camera System Design

The parallel-axis and convergent-axis systems are both formulated in the context of dual camera systems. While this encompasses as large body of the work in active systems, it leaves a large opening for extending this theory to the design of recovery systems for N camera systems. Also, the technique uses a measure of optimality that averages the error over the entire workspace. If more generic optimization constraints are incorporated, the richness of the theory would be significantly extended.

A.2 Integration of Representation into the Acquisition

In each of systems presented, the primary form of acquisition is three-dimensional depth information. An additional and costly processing step is required to process this

data into the TODEM representation. With some research effort the latency induced by this conversion would drastically improve the real-time performance of a TODEM design. This might be accomplished with special purpose hardware or efficient software that given a sequence of N input images will return an TODEM representation of the scene.

A.3 Extension of the Calibration Model

The TOPIC calibration module is of limited utility to vision problems in general since it is primarily aimed at the recovery paradigm. The calibration module could be integrated with representation and a user selected foundation algorithm if the TOPIC framework was extended.

A.4 Use of Generalized Representations

Elliptical Fourier Representations used in a Bayesian framework have some specific computational advantages for the class of problems that we have studied. In particular the construction of likelihood estimators is particularly simple in two-dimensional imagery. It would be nice to extend these types of schema to three dimensional using more advanced representations, such as Harmonic Sphereodial representations that have advanced three-dimensional properties. Also, a multi-dimensional representation would be more appropriate for the encoding of task dependent information.

A.5 Use of Composite Bayes Networks

Using linear combinations of parameter estimators has proven advantages in the computation of maximum a posterior networks. It would be of great use to extend the BAND network equations to include functional composites of principal model estimators. Not only would this enhance network performance, but it would make more efficient use of the available data.

VITA

Scott Mason Thayer was born in Brookings, South Dakota in April 1969. He was awarded the Bachelor of Science degree in Electrical Engineering by the University of Tennessee, Knoxville in May of 1991. In August of 1993, he received the Master of Science degree in Electrical Engineering, also from UTK. He was awarded the Doctor of Philosophy degree in Electrical Engineering from the University of Tennessee in May of 1998. Currently, he is employed by the Field Robotics Center of Carnegie Mellon University where his interests include, sensor-based robotics, real-time image processing, optimization in vision algorithms, and application of computer vision theory to field robotics.