Vergence control for a biologically inspired binocular active vision system

Zhang, Xuejie

2010

Zhang, X. (2010). Vergence control for a biologically inspired binocular active vision system.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/50632>

<https://doi.org/10.32657/10356/50632>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**VERGENCE CONTROL FOR A BIOLOGICALLY INSPIRED
BINOCULAR ACTIVE VISION SYSTEM**

ZHANG XUEJIE

SCHOOL OF COMPUTER ENGINEERING

2010

VERGENCE CONTROL FOR A BIOLOGICALLY INSPIRED BINOCULAR ACTIVE VISION SYSTEM

ZHANG XUEJIE

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2010

Acknowledgement

I would like to express my thanks to the School of Computer Engineering, Nanyang Technological University for giving me the opportunity to pursue my PhD study.

I would like to specially thank my supervisor, Asst Prof Tay Leng Phuan, Alex, for his encouragement, guidance and support during my study. He is always reachable and willing to discuss and share thoughts in both research and life. His attitude, passions and expertise helped me in all the time of my research and thesis writing. Without his guidance, this thesis would not have been completed.

I would also like to thank Assoc Prof Lin Weisi for his valuable suggestions to my research work. My thanks also go to my friends and colleagues in Emerging Research Lab for their support and suggestions: Li Dongtao, Song Hengjie, Jiang Xing, Cai Yundong, Wong Kia Yan, Nguyen Vu Anh, Dewi Muharyani Cendrawasih, Tse Rina, Thin Nandar Soe, and other students and staff in the lab.

My thanks also go to my thesis examiners for their comments and suggestions which greatly improved this thesis.

Lastly, I would like to express my sincere gratitude to my wife Xiu and my family for their love, patience, encouragement and sacrifice that they gave me to pursue my study. Without any of you, this would not be possible.

Table of Contents

ACKNOWLEDGEMENT	1
TABLE OF CONTENTS	2
LIST OF FIGURES	6
LIST OF TABLES	11
ABSTRACT	12
1 INTRODUCTION	14
2 BACKGROUND AND RELATED WORK.....	18
2.1 THE HUMAN VISION SYSTEM	19
2.1.1 <i>The visual sensory system.....</i>	19
2.1.2 <i>Eye movements.....</i>	26
2.2 FOVEATING ACTIVE VISION SYSTEMS	31
2.2.1 <i>Saccade generation</i>	32
2.2.2 <i>Vergence control.....</i>	36
2.3 THE FLANN AND KFLANN NEURAL NETWORK	51
2.3.1 <i>The Fast Learning Artificial Neural Network.....</i>	51
2.3.2 <i>The K-iterations Fast Learning Artificial Neural Network</i>	53
3 THE COGV SYSTEM	55
3.1 AN OVERVIEW OF THE COGV SYSTEM	56
3.2 SACCADE	59
3.2.1 <i>The watershed model for saccade generation.....</i>	60
3.2.2 <i>The FLANN segmentation model for saccade generation.....</i>	63
3.2.3 <i>Saccade integration</i>	73
3.3 NEURAL CLASSIFICATION OF OBJECTS BASED ON GABOR SIGNATURE	74
3.3.1 <i>Gabor filters</i>	76
3.3.2 <i>Edge region and orientation retrieval.....</i>	79
3.3.3 <i>Gabor signatures</i>	82
3.3.4 <i>Classification of 2D objects using Gabor signatures</i>	84

3.3.5	<i>Discussions</i>	87
3.4	VERGENCE	88
3.4.1	<i>Vergence geometry</i>	88
3.4.2	<i>Vertical disparity and tonic vergence</i>	90
3.4.3	<i>Developed vergence control models</i>	91
4	TYPE-A VERGENCE: A LOG POLAR VERGENCE CONTROL SYSTEM	93
4.1	LOG POLAR TRANSFORMATION	94
4.2	CENTRAL VISUAL PATHWAY AND OCULAR DOMINANCE COLUMNS	97
4.3	THE VERGENCE CONTROL MODEL	98
4.3.1	<i>The algorithm</i>	98
4.3.2	<i>Dynamics of disparity estimation</i>	101
4.4	EXPERIMENTAL RESULTS	104
4.4.1	<i>Comparisons with Cartesian or visual space</i>	104
4.4.2	<i>Repeatability of positioning</i>	108
4.4.3	<i>Depth and distance estimation</i>	109
4.5	DISCUSSIONS	110
5	TYPE-B VERGENCE: VERGENCE CONTROL USING DISPARITY ENERGY NEURONS	112
5.1	THE DISPARITY ENERGY MODEL.....	113
5.1.1	<i>Gabor filters and disparity energy cells</i>	114
5.1.2	<i>The position-shift model and the phase-shift model</i>	116
5.1.3	<i>Pooling and population responses</i>	118
5.2	THE PYRAMIDAL MODEL FOR DISPARITY AND VERGENCE	121
5.2.1	<i>Orientation and spatial pooling</i>	123
5.2.2	<i>Multilevel disparity estimation</i>	126
5.2.3	<i>Functional equivalence</i>	128
5.2.4	<i>Vergence control</i>	129
5.3	EXPERIMENTAL RESULTS	132
5.3.1	<i>Vergence control and depth estimation</i>	132
5.3.2	<i>Vertical disparity minimization</i>	136

5.3.3	<i>Accuracy and computational complexity of disparity estimation.....</i>	137
5.3.4	<i>Comparison with other related methods.....</i>	139
5.4	DISCUSSIONS.....	144
6	TYPE-C VERGENCE: A PYRAMIDAL CARTESIAN CORRELATION MODEL FOR VERGENCE CONTROL	145
6.1	THE PCC VERGENCE CONTROL MODEL.....	146
6.2	DISPARITY ESTIMATION FOR VERGENCE CONTROL.....	149
6.3	EXPERIMENTAL RESULTS	151
6.3.1	<i>Combining saccade and vergence: exploration of the scene</i>	151
6.3.2	<i>The zero-disparity region</i>	152
6.3.3	<i>Depth estimation</i>	153
6.4	THE PCAD VERGENCE CONTROL MODEL: EXTENSION TO COLOR IMAGES WITH COLOR MASKING.....	153
6.5	COMPARISON WITH EXISTING METHODS.....	157
6.6	DISCUSSIONS.....	158
7	TYPE-D VERGENCE: A PYRAMIDAL LOG POLAR CORRELATION MODEL FOR VERGENCE CONTROL	159
7.1	PLP: THE PYRAMIDAL LOG POLAR CORRELATION MODEL.....	160
7.1.1	<i>Normalized cross correlation in log polar domain</i>	160
7.1.2	<i>An evaluation of disparity measures and image space</i>	163
7.1.3	<i>Robust vergence in log polar domain</i>	165
7.1.4	<i>Coarse to fine disparity estimation using image pyramid</i>	167
7.2	EXPERIMENTAL RESULTS	171
7.2.1	<i>Vergence control.....</i>	171
7.2.2	<i>Tolerance to focusing and brightness</i>	176
7.2.3	<i>Depth estimation</i>	178
7.2.4	<i>Real time performance</i>	182
7.2.5	<i>Combining saccade and vergence: exploration of the scene</i>	183
7.2.6	<i>Comparison with existing log polar vergence control model.....</i>	185
7.3	DISCUSSIONS.....	186
8	CONCLUSIONS AND FUTURE WORK	188
8.1	SUMMARY	188

8.2 POSSIBLE FUTURE WORK	190
REFERENCES	192
INDEX	203
APPENDICES	204
APPENDIX A. THE DERIVATION OF PHASE SHIFT DISPARITY, IN ASSOCIATION WITH THE WORK OF CHEN AND QIAN [54].....	204
APPENDIX B. LIST OF PUBLICATIONS	208

List of Figures

FIGURE 1 THE EYEBALL	20
FIGURE 2 A SIMPLIFIED TOP VIEW SKETCH OF THE VISUAL PATHWAY	21
FIGURE 3 OCULAR DOMINANCE COLUMNS IN THE VISUAL CORTEX	22
FIGURE 4 CENTER-SURROUND RECEPTIVE FIELDS	23
FIGURE 5 RECEPTIVE FIELDS OF SIMPLE CELLS	24
FIGURE 6 RECEPTIVE FIELDS OF COMPLEX CELLS	24
FIGURE 7 THE HIERARCHY OF RECEPTIVE FIELDS FROM GANGLION CELLS TO SIMPLE CELLS AND TO COMPLEX CELLS, ADAPTED FROM [1].	25
FIGURE 8 AN ILLUSTRATION OF THE LOG POLAR MAPPING ON THE VISUAL PLANE. BLACK DOTS DENOTE CENTERS OF THE RECEPTIVE FIELDS.	26
FIGURE 9 PANUM'S FUSIONAL AREA	31
FIGURE 10 (A) THE ORIGINAL IMAGE AND (B) THE SALIENCY MAP	36
FIGURE 11 VERGENCE GEOMETRY	37
FIGURE 12 PARALLEL AND VERGING BINOCULAR SYSTEMS	38
FIGURE 13 A DISPARITY ENERGY UNIT	45
FIGURE 14 THE SACCAD-VERGENCE CYCLE: (A) MONOCULAR DETECTION OF TARGET, (B) SACCAD THROUGH PARALLEL VERSION MOVEMENT AND (C) VERGENCE MOVEMENT	50
FIGURE 15 THE CogV SYSTEM	56
FIGURE 16 THE FLOWCHART OF THE CogV SYSTEM	57
FIGURE 17 DEVELOPED MODELS FOR THE CogV SYSTEM	59
FIGURE 18 WATER FLOW IN WATERSHED MODEL	61
FIGURE 19 WATERSHED ON A REAL SCENE: (A) INPUT IMAGE, (B) EDGES, (C) WATER STRENGTH AND (D) OBJECT POSITIONS	62
FIGURE 20 RESISTANCE TO NOISE OF THE WATERSHED MODEL: (A) EDGES, (B) WATER STRENGTH AND (C) CELLS CROSSING THRESHOLD SHOW PRESENCE OF OBJECTS	63
FIGURE 21 THE SC MODEL FOR SACCAD GENERATION	63
FIGURE 22 THE FLANN SEGMENTATION MODEL	64
FIGURE 23 FLANN SEGMENTATION OF SCENE IMAGES: (A) ORIGINAL IMAGE, (B) FLANN SEGMENTATION WITH ENTROPY THRESHOLD, (C) ADD TILT CONSTRAINTS AND (D) ADD GABOR EDGE THRESHOLD	69

FIGURE 24 SACCADDE GENERATION BY SEGMENTATION AND ATTENTION: (A) ORIGINAL IMAGE, (B) FLANN SEGMENTATION, (C) ATTENTION MAP AND (D) CANDIDATE SACCADDES.....	71
FIGURE 25 SACCADDE GENERATED WITH ATTENTION THRESHOLD $T_A=0.4$, $T_B = 0.4$ AND $T_C = 0.2$	72
FIGURE 26 SIMULATED SACCADDE BY ACCEPTING MOUSE CLICKS FROM HUMAN SUBJECTS	72
FIGURE 27 THE LOCALLY CONSTRUCTED GLOBAL MOTOR MAP	73
FIGURE 28 ANTISYMMETRIC AND SYMMETRIC GABOR FILTERS.....	75
FIGURE 29 GABOR FILTER PROFILES WITH 4 SCALES AND 8 ORIENTATIONS, NORMALIZED TO [0, 1]: (A) ANTISYMMETRIC GABOR FILTERS AND (B) SYMMETRIC GABOR FILTERS	77
FIGURE 30 EDGE DETECTION BY GABOR FILTERS	77
FIGURE 31 EDGE ORIENTATION DETECTION USING GABOR FILTERS: (A) INPUT IMAGE, (B) ORIENTATION IMAGE BY ANTISYMMETRIC GABOR FILTERING AND (C) ORIENTATION IMAGE BY SYMMETRIC GABOR FILTERING.....	78
FIGURE 32 (A) LARGEASYM GABOR FILTER, (B) SMALLASYM GABOR FILTER, (C) INPUT IMAGE AND (D) HORIZONTAL PROFILES OF THE RESPONSES OF LARGEASYM AND SMALLASYM GABOR FILTER BANKS	80
FIGURE 33 EDGE ORIENTATION DETECTION.....	81
FIGURE 34 (A) EDGE ORIENTATION DETECTION USING ANTISYMMETRIC FOLLOWED BY SYMMETRIC GABOR FILTER BANK, (B) EDGE REGION RETRIEVAL AND (C) EDGE ORIENTATIONS	82
FIGURE 35 GABOR SIGNATURE GENERATION	82
FIGURE 36 SCALED AND ROTATED IMAGES AND THEIR GABOR SIGNATURES: (A) ORIGINAL IMAGES AND (B) IMAGES WITH GAUSSIAN NOISE.....	83
FIGURE 37 SAMPLE OBJECTS.....	84
FIGURE 38 DISPARITY AND VERGENCE.....	89
FIGURE 39 AN ILLUSTRATION OF THE RELATIONSHIP BETWEEN THE VISUAL SPACE AND THE LOG POLAR TRANSFORMED CORTICAL SPACE	95
FIGURE 40 LOG POLAR TRANSFORMATION: (A) ORIGINAL IMAGE AND (B) LOG POLAR IMAGE	96
FIGURE 41 CYCLICAL FLOWCHART FOR VERGENCE CONTROL BEGINNING WITH A VIEW CHANGE	99
FIGURE 42 ILLUSTRATION OF THE ALGORITHM.....	101
FIGURE 43 THE FLOWCHART OF THE ALGORITHM FOR COMPARISON	104
FIGURE 44 LEFT AND RIGHT IMAGES FOVEATING ON THE SAME OBJECT	105

FIGURE 45 VERGENCE ERROR ESTIMATION. VERTICAL AXIS SHOWS THE DISPARITY ESTIMATED AND HORIZONTAL AXIS SHOWS THE TRUE DISPARITY AT A STEP OF 2.5°	106
FIGURE 46 THE INTERLACED EDGE IMAGE. RED COLOR IS FROM THE LEFT EDGE IMAGE. GREEN COLOR IS FROM THE RIGHT EDGE IMAGE. OVERLAPPED EDGES ARE SHOWN IN YELLOW.	107
FIGURE 47 THE EDGE DETECTION ON LOG POLAR IMAGES WITH THRESHOLD (A) 10, (B) 20, (C) 40 AND (D) 80	108
FIGURE 48 VERGENCE CONTROL RESULTS: THE INITIAL PARALLEL IMAGE PAIR AND IMAGE PAIRS AFTER VERGENCE CONTROL ON AN OBJECT. THE DOTTED SQUARES SHOW THE FOVEA OF EACH IMAGE.	108
FIGURE 49 (A) CONFIGURATION OF THE EXPERIMENT: SIDE VIEW AND TOP VIEW (B) INITIAL STATUS AND VERGENCE ON THE OBJECTS	110
FIGURE 50 A COMPLEX CELL EXAMPLE USING THE PHASE SHIFT DISPARITY ENERGY MODEL: $\phi_L=0$, $\phi_R=\pi/2$, $\Omega=\pi/4$. THE PREFERRED DISPARITY IS +2. THE INPUT IMAGE SIZE IS 200×200 AND THE RECEPTIVE FIELD SIZE WAS 29×29 . THE CENTER OF THE BINOCULAR IMAGE PAIR IS EXPECTED TO CONVERGE TO A DISPARITY OF 0.	118
FIGURE 51 THE ORIENTATION POOLING AND SPATIAL POOLING PROCESS.....	120
FIGURE 52 THE IMAGE PYRAMID AND FIXED TUNING SPATIAL FILTERS.....	122
FIGURE 53 (A) THE PYRAMIDAL METHOD WHERE THE RECEPTIVE FIELD SIZE IS FIXED AND THE IMAGES ARE SUB-SAMPLED TO FORM A PYRAMID AND (B) THE COARSE-TO-FINE METHOD [54] WHERE THE IMAGE DOES NOT CHANGE AND THE FILTERS ARE SCALED.....	123
FIGURE 54 RESPONSES OF A GROUP OF COMPLEX CELLS WITH ORIENTATION $\theta=\pi/4$	124
FIGURE 55 THE ORIENTATION POOLING	125
FIGURE 56 THE FLOWCHART OF THE PYRAMIDAL DISPARITY ENERGY MODEL	126
FIGURE 57 DISPARITY ESTIMATION RESULTS ON A 200×200 RANDOM DOT STEREOGRAM WITH DOT DENSITY 0.5. THE LAST COLUMN SHOWS A MESH PLOT OF THE DISPARITY MAP GENERATED. LEVEL 1 AND LEVEL 2 RESULTS OF THE PYRAMIDAL MODEL WAS $\frac{1}{4}$ AND $\frac{1}{2}$ SIZES OF LEVEL 3 AND WERE SCALED TO THE LEVEL 3 SIZE FOR COMPARISON PURPOSE.....	128
FIGURE 58 THE FLOWCHART OF VERGENCE CONTROL.....	131
FIGURE 59 THE VERGENCE CONTROL RESULTS: BINOCULAR IMAGES ARE SHOWN IN PAIRS AND THE GRIDDED BOX INDICATES THE CENTER OF IMAGE	134
FIGURE 60 FAILURE CASES OF VERGENCE CONTROL.....	136
FIGURE 61 (A) VERGENCE CONTROL WITH VERTICAL DISPARITY CORRECTION AND (B) VERGENCE CONTROL WITHOUT VERTICAL DISPARITY CORRECTION	136

FIGURE 62 THE VERGENCE CONTROL PROCESS. THE FRAME NUMBERS JUST INDICATE THE SUCCESSIVE SEQUENCE OF IMAGES, NOT THE TRUE VIDEO FRAME NUMBER	137
FIGURE 63 VERGENCE CONTROL RESULTS BY STURZL'S METHOD [47]: (A) VERGENCE CONTROL RESULTS, (B) A WRONG VERGENCE AND (C) A VERTICAL MISALIGNMENT (WINDOW CROPPED FROM IMAGE CENTER).....	142
FIGURE 64 (A) THE INITIAL CAMERA POSITION BEFORE VERGENCE CONTROL FOR THE VERGENCE RESULT IN FIGURE 63B, (B) THE CORRECT ENERGY PEAK AT DISPARITY -7 WAS SURPASSED BY ANOTHER PEAK AT DISPARITY +40	143
FIGURE 65 THE TESTING SCENE	143
FIGURE 66 IMAGE PYRAMID AND DISPARITY ESTIMATION.....	147
FIGURE 67 INITIAL VERGENCE: BLACK CROSSES INDICATE THE CENTERS OF THE TWO IMAGES.	150
FIGURE 68 (A) THE DISPARITY ESTIMATION CURVE AND (B) THE CURVE SUPERIMPOSED WITH THE LEFT IMAGE	150
FIGURE 69 EXPLORATION OF THE SCENE: THE FOVEA IMAGES IN THE PROCESS OF SACCADDE AND VERGENCE.	151
FIGURE 70 SEGMENTATION OF OBJECTS BASED ON DISPARITY: (A) REGIONS BASED SOLELY ON DISPARITY AND (B) REGIONS AFTER MORPHOLOGICAL OPERATIONS.	152
FIGURE 71 THE IMAGE PATCHES AND THE GENERATED MASKS	155
FIGURE 72 (A) VERGENCE CONTROL WITH THE PCC MODEL AND (B) VERGENCE CONTROL WITH THE PCAD MODEL.....	156
FIGURE 73 AN IMAGE AND ITS COLOR MASK	157
FIGURE 74 THE BINOCULAR IMAGE PAIR.....	161
FIGURE 75 DISPARITY TUNING CURVES WITH LOG POLAR AND CARTESIAN IMAGES: (A) VARYING LOG POLAR IMAGE RESOLUTION, (B) VARYING CARTESIAN IMAGE AND TRANSFORM TO FIXED SIZE LOG POLAR IMAGES AND (C) VARYING WINDOW SIZES IN CARTESIAN DOMAIN.	162
FIGURE 76 THE LEFT IMAGE OF THE EXPERIMENTS. THE FOREGROUND OBJECT WIDTH IS (A) 50% (B) 25% AND (C) 12.5% OF THE FULL IMAGE WIDTH.	163
FIGURE 77 DISPARITY TUNING CURVES FOR SAD, SSD, NCC AND NCC IN LOG POLAR SPACE. THE SOLID VERTICAL LINE REFERS TO THE FOREGROUND, AND THE DOTTED VERTICAL LINE REFERS TO THE WALL AND DOOR BEHIND.	164
FIGURE 78 THE LOG POLAR TRANSFORMATION IN THE IMAGE PYRAMID	167
FIGURE 79 DISPARITY ESTIMATION USING LOG POLAR IMAGE: BLACK PIXEL DENOTES THE REFERENCE POSITION IN THE LEFT IMAGE AND THE CANDIDATE POSITIONS IN THE RIGHT IMAGE. (A) TRANSFORMATION OF THE LEFT IMAGE WITH THE REFERENCE POSITION AS ORIGIN AND (B) TRANSFORMATION OF THE RIGHT IMAGE WITH THE CANDIDATE POSITIONS AS ORIGINS ...	169
FIGURE 80 VERGENCE CONTROL RESULTS	172

FIGURE 81 VERGENCE CONTROL RESULTS WHEN THE TWO CAMERAS WERE SET TO DIFFERENT BRIGHTNESS.....	173
FIGURE 82 VERGENCE CONTROL RESULTS IN A CORRIDOR SCENE	174
FIGURE 83 VERGENCE CONTROL WHEN THE TARGET IS REMOVED: (A) BINOCULAR IMAGES AT THE BEGINNING OF VERGENCE CONTROL AND (B) FOVEA REGION SEQUENCE UNDER VERGENCE CONTROL.	175
FIGURE 84 VERGENCE CONTROL WHEN THE TARGET IS REMOVED: (A) VERGENCE ANGLES AND (B) ESTIMATED DEPTH	176
FIGURE 85 THE LEFT CAMERA WAS ADJUSTED TO 7 LEVELS OF FOCUSING	176
FIGURE 86 THE LEFT CAMERA WAS ADJUSTED TO 7 LEVELS OF BRIGHTNESS	177
FIGURE 87 THE EXPERIMENTAL ENVIRONMENT: (A) A CAMERA VIEW AND (B) THE PLACEMENT OF OBJECTS. NUMBERS NEAR THE OBJECT INDICATES THE HEIGHT OF THE OBJECT.....	179
FIGURE 88 (A, D) SACCADIC MOVEMENT BY THE MASTER CAMERA, (B, E) SACCADIC AND VERGENCE MOVEMENTS BY THE SLAVE CAMERA AND (C, F) RECONSTRUCTION OF THE POSITIONS OF FIXATED OBJECTS. CAMERAS WERE SHOWN AS TRIANGLES AND OBJECTS WERE SHOWN AS CIRCLES.	184

List of Tables

TABLE 1 VERGENCE CONTROL SYSTEMS	40
TABLE 2 PARAMETER SETTINGS OF GABOR FILTER BANKS	79
TABLE 3 CLUSTERING OF GABOR SIGNATURES	86
TABLE 4 CLASSIFICATION OF GABOR SIGNATURES.....	87
TABLE 5 VERGENCE CONTROL ERRORS	109
TABLE 6 DEPTH AND DISTANCE ESTIMATION.....	110
TABLE 7 PARAMETERS OF THE PROPOSED PYRAMIDAL MODEL AND THE COARSE-TO-FINE DISPARITY ENERGY MODEL [54]	127
TABLE 8 DEPTH ESTIMATION RESULTS	135
TABLE 9 THE ACCURACY OF DISPARITY ESTIMATION FOR THE 200×200 RANDOM DOT STEREOGRAM.....	138
TABLE 10 THE PROCESSING TIME OF THE PYRAMIDAL MODEL AND THE COARSE-TO-FINE MODEL	139
TABLE 11 DEPTH ESTIMATION RESULTS	144
TABLE 12 DEPTH ESTIMATION RESULTS	153
TABLE 13 DEPTH ESTIMATION RESULTS	158
TABLE 14 DEPTH ESTIMATION UNDER DIFFERENT FOCUSING.....	177
TABLE 15 DEPTH ESTIMATION UNDER DIFFERENT BRIGHTNESS	178
TABLE 16 DEPTH ESTIMATION RESULTS OF THE PLP MODEL	179
TABLE 17 MODELS IN THE DEPTH ESTIMATION EXPERIMENT	180
TABLE 18 DEPTH ESTIMATION ERROR AND STANDARD DEVIATION OF ERROR.....	181
TABLE 19 PROCESSING TIME OF THE MODELS	183

Abstract

The human naturally possesses a robust and effective binocular vision system that utilizes saccade and vergence eye movements to explore the visual environment. This thesis studies how low-level visual functions such as vergence and saccade can provide a basis to build an active vision system that is able to robustly support scene exploration. Several available vision techniques were examined for suitability of application into human-like binocular active vision systems. A biologically inspired binocular active vision proof-of-concept platform was designed and built.

This thesis focuses on the development of biologically inspired vergence control models. Vergence is the capability of human vision to gaze two eyes on a single spot in space. Through the study on the human vision system, several vergence control models were developed for the binocular vision platform. An initial vergence control model based on edge features in log polar space was first developed to achieve vergence in a simplified environment. Through further research and experimentation, investigations progressed onto a pyramidal vergence control model using disparity energy neurons. This second model utilizes fixed tuning disparity energy neurons and a pyramidal image structure to estimate disparity, resulting in near real time vergence control. Subsequently a pyramidal area correlation model was developed for simple vergence capabilities. As the search for a better solution prevailed, the model evolved into a disparity estimation model using a pyramidal fusion of log polar images. This final model superseded the previous models and exhibited robustness, providing a fundamental basis for developing real world applications for binocular foveation without object registration techniques or very accurate camera lens calibrations. Besides the developed vergence control models,

the breadth of supporting technologies included the building of mechanisms that simulated saccadic eye movements through the use of image segmentation and visual attention approaches.

Through the study and development of the biologically inspired components, the developed binocular vision system was able to select fixation points, reconstruct depth of the attended positions through vergence and build a simple fixation based spatial understanding of the environment. This thesis is presented with a balance of research rigour in the vergence domain and a broad-based understanding of human visual concepts dealing with viable computational solutions to saccade and vergence.

1 Introduction

The world can be perceived through the five senses: vision, hearing, taste, smell and touch. Amongst these, the vision system is one of the most important sensory systems for perception, action and learning. About 15% of the primate cortex is dedicated for the vision system [1]. The human vision system is a robust binocular system for sensing and processing visual information that is extremely quick to appreciate the environment. It is complex, consisting of visual sensory, motor, visual transmission and visual processing components. The eyeballs are controlled by six extra ocular muscles that respond to neural signals, empowering the eyes with saccadic burst-speeds of up to $800^\circ/\text{sec}$ [2]. These integral components make the human vision both a swift and flexible system.

Recent interest in active vision systems for autonomous robots has led researchers to explore studies in biological research in search of applicable techniques in computational vision. Researchers have been simulating various aspects of the human vision system such as retina-like sensors [3-4], the control and stabilization of eye movements [5-6], and the neural processing of visual information [7-8]. The humanoid robot has also become a research area that has attracted much attention.

This thesis explores the breadth of visual understanding for a cognitive autonomous vehicle (CAV) project and focuses primarily on bio-inspired computational vergence control. The context of the ‘cognitive’ is in the functional infrastructure where the

system is capable of supporting the creation of basic primitives from the vehicle's environment thereby enabling higher level intelligent processes. In the project domain specified, key information such as objects of interest, relative distances between objects and traversable paths are crucial to the CAV's survivability. These tasks are achieved through a functional framework that combines computationally derived attention, saccade generation, vergence control, target acquisition and identification. While the broad spectrum of the framework is built upon various facets of available technologies, these technologies function as supporting ligaments for an in-depth study of bio-inspired vergence. The focus of this thesis lies rooted in the design of a vergence system that complements the visual framework and the objective is to develop robust vergence control models which can survive complex and cluttered environments.

Vergence is defined as the concurrent movement of the two eyes in opposing directions that provide convergence onto a single attended position in space. This is a natural capability of the human vision system to fixate both eyes on a single object, obtaining two perspective images at the early vision pathways which eventually fuse as a stereo image when signals reach the higher levels of the cortex. Also residing within the visual system is a uniquely structured retina where the fovea or center of retina possesses the highest resolution and acuity. This variable acuity retinal structure provides a unique characteristic for image acquisition and is consequently complemented with a complex active gaze control system that voluntarily shifts the gaze to a point of attention. With vergence, the attended region is empowered with high quality stereo details that are collated over a span of several saccades. The stitched sequence of these saccades provides a positional reference frame that is fundamental to scene understanding. Apart from the physical attributes used in the comparisons, considerations from the cognitive

psychology perspective are also introduced. Humans in active conversation project an attention of focus by fixating their eyes to participants of the conversation. This behaviour is perceived by other parties who are participating in the conversation as a protocol for engaging in discussions. This important interactive process gives rise to the importance of vergence, raising the importance for a robot's ability to connect in human interaction, providing an engaging experience.

Vergence control in a binocular active vision system aims to direct the two cameras towards the same point in space. In active vision systems, problems in simulating vergence eye movements include: What cues does the biological vision system provide for computational vergence? Is it possible to use these cues for the development of algorithms in active vision systems and robot applications? What are the difficulties in vergence control? What are the important factors affecting the performance of a vergence control model and how is performance measured? This thesis attempts to answer these questions through the investigation into the biological vision system, the development of new vergence control algorithms, the analysis of the working environment and the system design. The purpose of this work is to design and implement a binocular vision system that has robust vergence and simple saccade capabilities. The investigation in this thesis ventures into various strategies for vergence control and makes comparisons across a wide spectrum of proposed systems. In the course of these investigations, a pyramidal log polar vergence algorithm that provides a robust and real time performance for vergence control in complex scenes is developed. A thorough comparison is done between this model and other available models.

This thesis can broadly be divided into four parts. Chapter 2 presents the background and related work, encompassing the basis of the human vision system, active vision systems and explores the details of existing vision techniques applicable to vergence. Chapter 3 describes the hardware and architectural design of the proposed cognitive vision system, namely CogV. Image segmentation models for saccade generation and a Gabor signature based object recognition model are also presented in this chapter. Chapter 4-7 provide detailed elaboration of the vergence control models and consolidate the comparisons of available methods. Chapter 8 provides the conclusions to the investigation and presents possible future work.

2 Background and Related Work

The human vision system is a complex system consisting of many individual components that can be modeled for robust vision [9]. It involves many levels of neuronal processes that branch and extend from the retinal cells to the frontal cortex. This complex network is further augmented by the many feedback mechanisms of inter-cortical and intra-cortical signals in the brain cortex. The result of these interactions is the human's ability for complex perceptual capabilities. The simplest visual task is still a complex combination of neuronal processes and actuations involving many faculties in the brain. Vision research is a significant topic in psychology and neuroscience studies where the internal functionalities of the primate vision system are observed through psychology and physiology experimentations carried out on animals such as cats and monkeys. These studies provide crucial clues to the connection structures within the sensory system, the motor system, and the processing behaviour in the cortex.

During the recent years, active vision systems have attracted considerable interest in research and applications and have led to the development of verging stereo systems bearing similarity to the human vision. Active vision systems attempt to mimic primate biological vision with motorized cameras emulating the functions of the extra ocular muscles that control the movements of the eyes. Unlike passive or static vision systems, active vision systems physically rotate the cameras laterally to change the view point, thereby possessing the capability to obtain richer visual information from the

environment. Active vision systems often borrow principles from the biological system's geometrical design and have been able to exhibit robust and adaptive performance in the real world environment. Such biologically inspired designs aim to simulate how the human vision system works in a complex environment. Active vision research encompasses a wide range of topics including simulation of eye movements such as saccade, smooth pursuit, vergence control and motion tracking, and simulation of visual perception such as visual attention, stereopsis and visual navigation. This chapter provides a brief introduction to the human vision system, eye movements and active vision systems with vergence capability in the literature.

2.1 The human vision system

The human vision system consists of the sensory system (the eyeball), the motor system (the extra ocular and inner ocular muscles), the transmission system (the optic tract) and the processing system (the brain cortex). Studies on primate vision systems have been useful as these animals possess similar visual faculties as in the human vision system.

2.1.1 The visual sensory system

2.1.1.1 The eye and retina

Figure 1 shows a simple anatomical representation of the eyeball. Visual information enters the eye through the cornea, the iris, the lens and falls on the retina. The projected image on the retina is an inverted image of the visual field. The image is transmitted through optic nerves to the later part of the visual pathway.

Figure 1 The eyeball

There are two types of light receptors in the retina, the rods and the cones. The rods are responsible for our vision in dim light and the cones are responsible for our ability to see details and envision color [1]. Although there are more rods in the retina than cones, the small area at the center of the retina are populated solely with a very high concentration of cones [10]. This area providing high visual acuity is called the fovea. The compactness of the cones provides the brain with high resolution and quality color images. An interesting feature of the retina is that the rods and cones lie beneath several cellular layers including the ganglion layer, plexiform layers and nuclear layers. The signal from the rods and the cones then goes back through several layers of bipolar cells, amacrine cells and horizontal cells before reaching the ganglion cells at the superficial layer of the retina. The ganglion cells send the visual signal to the midbrain and thalamus through the optic nerves.

2.1.1.2 The visual pathway

Anatomical studies of the vision system indicate the presence of a clear visual pathway from the retina to the brain. Signals exiting the eyes are transported through the optical nerve fibers to two lateral features known as Lateral Geniculate Nucleus (LGN). This feature found in the thalamus situated in the middle of the brain receives the signal and transmits the information to the primary visual cortex at the back of the brain. As a

result of the bi-hemispheric structure of the brain, it is not surprising that the visual field is also divided into two.

Figure 2 A simplified top view sketch of the visual pathway

As shown in Figure 2, the optic fibres from the nasal halves of the two retinas cross at the optic chiasma prior to the LGN in the visual pathway. The optic nerve fibres from the temporal halves of the two retinas do not cross and go to each half's ipsilateral (on the same side of the body) portion of the LGN. The LGN then transmits the visual information to the primary visual cortex through the optic tract. As a result, the primary visual cortex receives the visual information from the ipsilateral half of the two retinas, which is the contralateral half (on the opposite side) visual field of the two eyes. In addition to the LGN, the optic tract also feeds other parts of the brain including the superior colliculus (SC) which is involved in the control of eye movements.

An interesting phenomenon is the synergistic arrangement of the optic fibers as it reaches the primary visual cortex. As shown in Figure 3, the visual information from the two eyes is interlaced together to form a striped pattern of alternating eye signals. This striped pattern of information interlacing is known as ocular dominance columns [11] and is distinct in the primary visual cortex [1]. The neurons are organized to spatially preserve information within a local region so that localized binocular processing can be performed within the localized receptive fields. The presence of these stripes of alternating dominance prompted an investigation of the possibility that this naturally occurring organization is instrumental for the vision tasks. This is an important cue for binocular vision tasks such as binocular fusion and vergence control. Any disparity in the scene when viewed from the binocular system would usually results in a localized displacement seen within neighboring ocular dominance columns. It has also been discovered that the width of ocular dominance columns determines the stereo fusion limit [12].

Figure 3 Ocular dominance columns in the visual cortex

The visual information processing in the brain is far more complex than the transmission of visual signals. A simple vision task can be the result of orchestrated efforts of several parts of the brain, involving many stages of processing. The approach taken in this thesis requires a buildup of concepts beginning with the study of the human vision system. The visual functionalities were simulated using both computational and bio-inspired models.

2.1.1.3 Receptive fields

The receptive field of a neuron in the vision system is the region of the retina that excites this neuron [9]. This definition was also extended to include the specific properties of the stimulus that evoked the strongest response of the neuron [9]. The retina ganglion cells have on-center and off-center receptive fields, shown in Figure 4. The + and – sign denotes the excitation and inhibition areas in the receptive field. For on-center receptive field, a bright dot stimulus at the center of the receptive field excites the cell, causing it to fire. For off-center receptive field, a dark dot stimulus at the center of the receptive field causes the cell to fire. The LGN cells have similar properties with the ganglion cells. Hence, human vision tends to be relative where the bright or dark perception is not fully determined by the absolute strength of the visual signal. A dark dot may be perceived as bright if it is put onto a darker background and a bright dot may be perceived as dark when placed within a brighter context.

Figure 4 Center-surround receptive fields

The cells in the primary visual cortex show more complex characteristics resembling organized grouping of the signals from the LGN cells. There are different types of cells in the primary visual cortex. These include the simple cells and complex cells. Simple cells respond to oriented stimulus such as edges and lines, shown in Figure 5. The receptive fields of simple cells can be divided into excitatory and inhibitory regions. A proper oriented line or edge stimulus at the center of the receptive field will excite the simple cell.

Figure 5 Receptive fields of simple cells

Complex cells are the next stage processing. They tend to have larger receptive fields. A properly oriented line stimulus within the receptive field will cause the cell to fire, no matter where it is placed in the receptive field, as shown in Figure 6. There are also other types of complex cells with more complex receptive fields and receptive properties. For example, some complex cells have binocular receptive fields exhibiting preference to binocular disparities [13].

Figure 6 Receptive fields of complex cells

It is still not clear how the complex cells are built up [1] while there are a fruitful number of models proposed to model how the cells work. Hubel et al through a systematic study of retinal cells proposed a possible way of constructing the simple cells and complex cells using a hierarchical structure of cells [13], shown in Figure 7.

Figure 7 The hierarchy of receptive fields from ganglion cells to simple cells and to complex cells, adapted from [1].

2.1.1.4 The retino-cortical map

It was discovered that the visual information in the visual pathway is with a well organized structure. The connections from the eyes to the LGN and from LGN to the primary visual cortex were found to be topographically synchronized. There is a precise correspondence between a given location in the primary visual cortex and that in the subjective visual field. The retinotopic mapping of the visual field to the surface of the primary visual cortex can in fact be characterized as a logarithmic conformal mapping [14-19]. The architecture is such that a large portion of the primary visual cortex is mapped to the small, central portion of the visual field. This mapping was found to exhibit a log polar topology and the over representation of the fovea at the visual cortex is known as cortical magnification. The log polar transformation is summarized in Equations (1) and (2).

$$\rho = \log_a \sqrt{(x - x_0)^2 + (y - y_0)^2} \quad (1)$$

$$\varphi = \arctan\left(\frac{y - y_0}{x - x_0}\right) \quad (2)$$

where (x_0, y_0) is the origin of the visual plane,

ρ and φ are the two axis's of the log polar domain

Figure 8 is an illustration of this log polar relationship between points in the cortical plane and those in the visual plane. The receptive field size in the visual plane becomes larger as the distance to the fovea increases. A compact arrangement of receptive cells is found at the center of the map. This provides for the high resolution in the fovea.

Figure 8 An illustration of the log polar mapping on the visual plane. Black dots denote centers of the receptive fields.

2.1.2 Eye movements

Stabilizing the retina with regard to the outside world and aligning the retina with stationary or moving targets is a critical challenge to effective vision [9]. The mechanisms that control eye movements in the human vision system can be divided into two principal classes: gaze stabilization mechanisms and gaze shifting mechanisms.

2.1.2.1 Gaze stabilization

If both the environment and eyes stay stationary, the image projected on our retina will always be stable and clear. However, the human vision system is active vision system and the environment is constantly changing. The gaze stabilization mechanisms in the human brain help to compensate self-motion and visual world shifting to stabilize the image projected on the retina. These mechanisms make the line of sight constant with respect to the environment, despite the eye, head and body movements. The two sub-classes of gaze stabilization mechanism are vestibulo-ocular systems and optokinetic systems [9]. The vestibulo-ocular system determine the rate of head rotation through the semi-circular canals and the optokinetic system computes the speed and direction of visual world shifting on the retina through the photoreceptors themselves [9]. These neural systems then compensate the rotation by a counter rotation of the eyes through actuating the extra ocular muscles. Due to the difference between the motion estimation sources, the vestibulo-ocular system works most efficiently in high velocity movements and the optokinetic system works most efficiently in low velocity movements.

2.1.2.2 Gaze shifting

Gaze shifting mechanisms only exist in vertebrates with foveal retina. Since high visual acuity resides at the fovea or center of the retina, gaze shifting allows the sporadic sequential attendance to areas of interest in high resolution and visual quality. Gaze shifting mechanisms can be divided into three classes: the saccadic system, the smooth pursuit system and the vergence system.

The saccadic system rapidly shifts gaze from one point to another. Saccade is the swift movement of both eyes simultaneously across a visual trajectory onto a new fixation point and provides the human with accurate visual positioning. Saccade provides us a way to focus on multiple discrete positions in the scene and selectively attend to important information in the environment. During and shortly before the saccade, the magnocellular pathway is selectively suppressed through a saccade masking or saccade suppression process [20]. In this very short millisecond-period of time, visual processing is selectively blocked so that vivid information is present before and after saccade. It is believed that the process of multiple saccadic observations reduces the processing load on the early visual areas and through the organized pathways, the brain is able to multiplex thought processes. The smooth pursuit system allows the fovea to track a moving target on a stationary background. The saccadic system and smooth pursuit system together are often referred to as versional systems [9].

The vergence control system exists only in a small subset of vertebrates possessing both foveal retina and binocular vision. Vergence directs both eyes to foveate singularly on a target, providing visual information for binocular perception. With the convergence and divergence of the gazes of two eyes, the depth information of the target can also be derived. Vergence is important for stereo vision, especially for reachable distance tasks such as reaching and grasping [21-22]. Without vergence, many of the primitive ocular functions of the human vision system, such as binocular fusion, will be impaired. There are four sources of vergence in the human vision system: binocular disparity, accommodation, tonic, and proximal vergence [9]. Binocular disparity is the displacement of visual positions stimulated by the same stimulus on the two retinas. This is a natural source for the vergence error signal. Accommodation status of the

lenses helps the derivation of object distance which can be used to estimate the motor error for vergence. Tonic is the default status of convergence in total darkness. Proximal vergence refers to the usage of cognitive depth cues to infer the distance to target and hence control vergence [9]. Through vergence, stereo perception on the foveated objects is enabled. Performing binocular synchronization requires the most basic operations of saccade and vergence. The combination of these eye movements allows the visual system to examine the scene at a fast speed through multiple fixation points [23].

In all, there are five basic mechanisms for eye movements: the vestibulo-ocular, optokinetic, saccadic, smooth pursuit, and vergence systems [9]. All human eye movements belong to one or more of these five classes. Although the five classes are different mechanisms residing in the brain, they each share portions of the same motor system. This implies that a single eye movement may involve several mechanisms. Thus, in order to simulate some of the complex human vision, it is necessary to appreciate and encapsulate portions of the visual system onto a single set of hardware and the cooperation of different mechanisms is important for robust vision.

2.1.2.3 Superior colliculus and eye movements

The superior colliculus is a laminated structure in the midbrain above the cerebral aqueduct [9]. While there is topographical visual organization in the visual cortex, neuroscience experiments showed that there is topographical organization of eye movements in the superior colliculus [24]. This was discovered when certain stimulus on the superior colliculus consistently resulted in an eye movement of certain amplitude and direction. The movement was also found to be relative with respect to the current

eye position. The neurons in the colliculus encode motor error rather than the absolute eye positions. Therefore the topographical organization of eye rotations in the colliculus takes reference from the fovea. When the saccadic system generates a saccadic eye movement from visual, auditory or somatosensory inputs, the neuron circuits in the brain compute the amplitude and direction of the desired movements and relay it to the colliculus to activate the corresponding eye movement.

2.1.2.4 Panum's fusional area

The human stereo perception has a restricted operating range related to the vergence state of the two eyes [25]. The human vision system has different levels of sensitivity for varying depths because of the finite resolution of retina and the unbounded depths in the scene. Panum's fusional area refers to the region of binocular single vision. If a point stimulus is on the retina of one eye, there is a region in the retina of the other eye where another point stimulus on the second eye can be binocularly fused with the first stimulus to provide stereo perception. Outside this region, the two stimuli cannot be fused. Therefore, the region in visual space corresponding to this retina area is Panum's fusional area, and objects in front and behind this region in the visual space will cause physiological diplopia or the double image phenomenon. For example, in Figure 9, object A is the fixation point. Objects A and B can be fused to provide stereoscopic vision but object C that is outside the Panum's fusional area cannot be fused.

Figure 9 Panum's fusional area.

Panum's fusional area depends on the fixation point and the convergence state of the two eyes. It is a region corresponding to the horizontal radial loci bearing a similar distance of the fixation point. Panum's fusional area is associated with the cortical magnification factor and the size of ocular dominance columns in the primary visual cortex [12]. Assuming that Panum's fusional area results from the fixed width of the ocular dominance columns, this area delimits the range of possible local depth estimates which may be computed for a given point of fixation [26]. This advocates that binocular fusion can only be successful when the scene is of a degree of fusional match. Therefore vergence control is a crucial component in human vision. This is the advantage of the verging binocular vision system where vergence directs the eyes to fixate on different depth in the scene and easier stereo fusion for each depth is achieved.

2.2 Foveating active vision systems

Many of the active vision systems were inspired from the psychology and physiology studies conducted on the primate vision system. It is recognized that although computer vision is also achievable in various degrees through other methodologies that have

surfaced over the course of the research community, the biological visual systems may still possess significant advantages and provide potential improvements. Biological components in the primate vision system have been used widely for developing robust computational vision models. One such example is the robust feature retrieval for object recognition, inspired from studies on the early visual cortex of the human vision system [8]. It is believed that the clue to designing an effective active vision system lies in the natural design of the primate vision system. This is also one of the reasons why researchers have been developing human-like vision systems.

In this thesis, the focus is on foveating or verging binocular vision systems. The definition of foveating binocular systems encompasses systems that actively pan and tilt in order to centrally locate the intended fixation point at the centers of both captured images. Foveating refers to the ability for the binocular system to converge at the same point in space. In this thesis, simulation is conducted to envisage vergence and saccade possibilities in the human vision system through computational vision models. This is believed to be both important and useful for many real world robot applications.

2.2.1 Saccade generation

Saccades are probably the fastest movements produced by the human body. While the mechanism for the control of saccadic movement has been well studied, very little about how saccadic eye movements are generated is known to us [9]. In studies on spatial visual attention, there exists theories that the saccade is related to the spatial attention [27]. The saliency based visual attention model has been considered one of the most biologically plausible human attention models [28]. This model creates visual saliency maps for different features such as color, intensity and orientation. Saliency is based on

the concept of center-surround contrast. The individual feature saliency maps were combined to form a visual attention map. Studies conducted indicate a partial correlation between this model with the actual human attention [29]. In artificial vision systems, saccade can be generated in different ways. Active vision systems such as [30-32] explore the environment using saccade based selective exploration or recognition. In this thesis, the saccade generation is simulated by the image segmentation models and color attention models. The objective is to retrieve informative positions in the visual scene. It is assumed that image segments correspond to smooth surfaces in the image and areas of interest are defined as proper sizes of segments with a certain level of attention. The following section is an introduction to Itti and Koch's saliency based visual attention model [28].

2.2.1.1 Itti and Koch's visual attention model

There are many visual attention models in the field of cognitive neural science and computer vision. The most widely used visual attention model is the visual attention model by Itti and Koch [28]. This model was built on a biologically plausible architecture by Koch and Ullman [33] and is related to the 'feature integration theory' which explains human visual search strategies [34]. It decomposes the visual input to several feature maps (intensity, color and orientation) and the visual information in these maps competes to provide individual saliency maps for each feature. The competition inside a feature map is through a center-surround difference approach realized by cross-scale subtraction using a multi-resolution feature pyramid. The feature saliency maps are integrated to generate a saliency map for rapid scene analysis. The model models the bottom-up visual attention of the primate vision system. The following section shows how a visual attention map is generated.

2.2.1.1.1 Generating the feature map pyramids

From the original *rgb* color images, intensity image *I* is formed as $I=(r+g+b)/3$. Then the *r*, *g* and *b* planes are normalized by the intensity *I*. Positions with *I* less than 0.1 of its maximum were cleared to 0 as hue variations are not perceivable at low luminance. Subsequently four broadly-tuned color channels were created from the normalized *rgb* planes by Equation (3). Negative values were set to 0.

$$\begin{aligned} R &= r - \frac{g+b}{2} \\ G &= g - \frac{r+b}{2} \\ B &= b - \frac{r+g}{2} \\ Y &= \frac{r+g}{2} - \frac{|r-g|}{2} - b \end{aligned} \tag{3}$$

The intensity and color pyramid is generated by progressively low-pass filtering and sub-sampling the original image [28]. The orientation pyramids are generated by filtering the intensity image by multi-scale oriented Gabor filters [35].

2.2.1.1.2 The individual saliency map for each feature

Five pyramidal strata representations were created for the plane *I*, *R*, *G*, *B*, and *Y*. Each pyramid consists of *N* levels with scaling ratio of 0.5 between successive levels (level 1 being the finest and level *N* the coarsest). In the implementation of [28], 640×480 color images and 9-level pyramids were used. Saliency is modeled as the center surround differences across different scales of the pyramid. Center surround differences are computed between a center finer scale *c* and a surround coarser scale *s* through a cross-scale subtraction. This is essentially a process of interpolating the coarser scale to finer scale and then conducting a point to point subtraction. The intensity saliency map is

constructed using Equation (4). Each center-surround difference map was normalized to the range [0, 1].

$$mapI = \sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(I(c, s)) \quad (4)$$

where $I(c, s) = |I(c) - I(s)|$

and $N(\)$ is the normalization operator

The color saliency is based on the idea of color opponency maps. The RG and GR opponency maps are created as $RG=R-G$ and $GR=G-R$. The BY and YB opponency maps are created as $BY=B-Y$ and $YB=Y-B$. The color saliency map is constructed through cross-scale subtraction of the RG , GR maps and BY , YB maps, shown in Equation (5).

$$mapC = \sum_{c=2}^4 \sum_{s=c+3}^{c+4} [N(mapRG(c, s)) + N(mapBY(c, s))] \quad (5)$$

where $mapRG(c, s) = |RG(c) - GR(s)|$

and $mapBY(c, s) = |BY(c) - YB(s)|$

The orientation saliency maps are generated similarly. 4 orientations were used.

$$mapO = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N \left(\sum_{c=2}^4 \sum_{s=c+3}^{c+4} N(mapO(c, s, \theta)) \right) \quad (6)$$

where $mapO(c, s, \theta) = |O(c, \theta) - O(s, \theta)|$

After this, the final saliency map is generated by combining all the feature saliency maps through cross-scale addition, which scales each map to the resolution of level 5 and sums them to one map.

$$attentionMap = \frac{1}{3} (mapI + mapC + mapO)$$

The final saliency map describes the bottom-up attention generated from the visual input. Figure 10 showed an example of the saliency map. Objects or regions that stand out from the background receive higher attention level.

Figure 10 (a) The original image and (b) the saliency map

2.2.2 Vergence control

Vergence control is an active process to centralize the two eyes (cameras) such that they foveate on a single point in space. It differs from the static or fixed-camera stereo vision system in that the static stereo vision system finds the disparity between corresponding points or reconstructs a 3D map, given the two projected images. However, the foveating binocular system can derive the 3D information of its attended position or region by direct computation through the motor configurations. In this case, the active binocular vision systems can potentially be more robust, removing the intricate reliance on lens parameters to determine depth information. This depth information is useful for active vision systems or robots in a dynamic environment for applications such as object manipulation and obstacle avoidance.

Figure 11 Vergence geometry

As shown in Figure 11, with the two eyes L and R verging on a target T , the depth to target T can be calculated according to Equation (7).

$$D = \frac{B}{\tan(\alpha) - \tan(\beta)} \quad (7)$$

where B is the baseline distance between L and R

Let C be the cyclopean eye of the vision system, which is an imaginary eye located middle way between the two eyes. The cyclopean eye is relevant to the visual direction perception [36]. In a verging system, the cyclopean distance from the cyclopean eye C to the target T can be calculated according to Equation (8).

$$CT = \sqrt{\left(\frac{B}{2} + D \tan(\beta)\right)^2 + D^2} \quad (8)$$

Foveated binocular systems can provide a more robust stereo understanding for the attended areas because the fovea on the retina provides the highest acuity and the vergence process can limit the disparity of the object under fixation in the attended foveal region to nearly zero, providing better binocular fusion of the attended target. Figure 12 shows a parallel and a verging binocular vision system. The verging system possesses more overlapping and the binocular images captured provide better binocular

fusion, especially at the fixation point. Another important implication of vergence in social robotics is the ability to project attention. From the psychological aspect, human body-language interactions are usually complemented with visual fixations that function as cognitive cues to project one's focus of attention. It is from these perspectives that we derive the emphasis to attain the most fundamental building block of fixations, the vergence control.

Figure 12 Parallel and verging binocular systems

In the human vision system, the sensory system is always actively moving and shifting the gaze positions. If an object attracts the attention of the observer, the vision system will issue a saccadic movement command to shift the two eyes directly to the interesting area. Then vergence control and foveated observation will be carried out. Adjustments of the focusing and gaze directions make the vergence and stereo fusion possible. In active vision systems, these components can be simulated independently or cooperatively. Vergence control makes the binocular system an ego-centric system. The binocular disparity, positioning and control are all based on the current eye gaze positions.

The typical objective of computational stereo vision is to reconstruct the 3D environment from its projections in 2D images. In parallel systems the optical axis of the

two cameras are parallel and fixed while in the verging systems, the two views are not fixed and the system attempts to fixate on a single point in 3D space. When an object is in close proximity to the cameras, the parallel system encounters more difficulty in fusing the binocular images while verging system can change the gaze position to allow the two cameras to focus on the object at close proximity, providing better stereo information. Binocular disparity and accommodation are the two sources that have been extensively studied in computer vision research. Most existing vergence control models in literature, such as [37-38], belong in the category of binocular disparity estimation. The disparity obtained between a pair of binocular images can be used as a motor differential signal to verge the cameras. This section reviews existing disparity based vergence control systems. These include area correlation methods [39-42], phase based methods [6, 37, 43-44], cepstral filtering methods [38, 45], spatial variant methods or log polar transformation methods [5, 46], and disparity energy models [47-49]. There are also other models that rely on image features and optical flows [50-51]. Many disparity estimation techniques also employ hierarchical or coarse-to-fine control strategy to refine the disparity estimation [25, 37-38, 52-54]. Table 1 shows a list of active vision systems with verging capability. The following sections review these vergence control models in details.

Table 1 Vergence control systems

	Vergence error estimation	Control
Ching et al [39]	Accommodative vergence followed by template and sub-template matching using normalized cross correlation	Focus and disparity based proportional control
Abbott and Ahuja [42, 55]	Accommodative vergence followed by correlation based disparity	Focus and disparity based proportional control
Yamato [40], Roca et al [41]	Correlation and SAD based matching	Proportional control
Siebert and Wilson [37]	Phase based correspondence in a foveated pyramidal image structure	Proportional control by sum of weighted multi-level disparity
Hansen and Sommer [56]	Phase based multi resolution disparity with confidence measure from amplitude of filter responses. Local maxima of the confidence map is used as saccade control	PD controller
Theimer and Mallot [43]	Coarse to fine disparity estimation based on phase correlation with confidence from binocular amplitude contrast	Proportional control by sum of weighted multi-level disparity
Marefat et al [6]	Segmentation and phase detection	
Olson and Coombs [57]	Cepstral filtering	PD control
Taylor et al [38]	Cepstral filtering in truncated Gaussian pyramids	Proportional control
Capurro et al [5]	Dynamic vergence by fusion index and divergence motion estimation in log polar images	Derivative of fusion index and divergence motion
Oshiro et al [58]	Zero disparity filtering using log polar edge images and image translation in Cartesian domain	Proportional control
Manzotti et al [46]	Log polar camera sensors, normalized correlation using log polar to Cartesian mapping	Proportional control
Bernardino and Victor [59]	Correlation in log polar images	Derivative and preprogrammed control
Sturzl et al [47]	Disparity estimation using a position-shift disparity energy model	Proportional control

	Vergence error estimation	Control
Gibaldi et al [48]	Vergence in a virtual environment using phase-shift disparity energy model	
Tsang et al [49]	Virtual vergence through a neuromorphic implementation of the disparity energy model	
Piater et al [60]	Symmetric matching of Gabor oriented feature vectors in a column structure of the non-uniform foveal image	Move both camera to the best matching column pair
Choi et al [61]	Detection and matching of salient landmarks in both cameras	Move both cameras to the selected landmarks
Batista et al [62-63]	Dynamic vergence by target motion estimation	Difference of retinal optical flow
Kim et al [64]	Disparity flux of disparity flows (change of disparity) in an attention window with logarithmic sampling	Disparity flux
Marfil et al [65]	Hierarchical segmentation of stereo pairs and corresponding segment pairs provides the disparity	Proportional control using disparity between segments
Yim and Bovid [52]	Coarse to fine searching using sign correlation on a sign pyramid generated from a Laplacian pyramid	Proportional control

2.2.2.1 Correlation based correspondence

An intuitive way of establishing correspondences is to use explicit matching of image patches. Area correlation methods are common methods for matching regions or patches in the image for vergence control [39-41, 52, 60]. Equation (9) shows several commonly used correlation measures.

$$\begin{aligned}
 NCC(f, g) &= \frac{1}{n-1} \sum_{x,y} \frac{(f(x,y) - \bar{f})(g(x,y) - \bar{g})}{\sigma_f \sigma_g} \\
 SAD(f, g) &= \sum_{x,y} |f(x,y) - g(x,y)| \\
 SSD(f, g) &= \sum_{x,y} (f(x,y) - g(x,y))^2
 \end{aligned} \tag{9}$$

Olson used normalized cross correlation (NCC) to conduct stereo correspondence search in verging binocular image pair [25]. Autocorrelation between left and right image patches was used to support reliable disparity estimation of images with sparse information. Sum of absolute difference (SAD) and sum of squared difference (SSD) were also used widely in disparity estimation [66] and vergence control [40-41]. The advantage of normalized cross correlation is that it has invariance to camera gain and offset variations [67]. For example, when the binocular images bear different brightness due to mismatched camera configurations, normalized cross correlation produces a better performance than the other two correlation measures. This makes it suitable for real world applications where the two cameras may not support accurate calibrations.

Common problems of area correlation based method are the selection of the optimal window size and the determination of the correlation peak when several peaks exist. The former problem is due to the non-uniform disparity distributions in a local window such as boundaries of objects and occluded regions. The second problem is due to confounding information in the local area. While there is no universal setting suitable for the various conditions it is possible to constrain the scope to be application specific. These settings can be deduced through logical analysis and empirical experiments. Ching et al used a hierarchical area dividing technique to carry out area correlation in a coarse to fine manner [39]. Capurro et al and Manzotti et al also used correlation in a log polar image domain to determine the disparity [5, 46]. Recently in local area based stereo reconstruction, adaptive support weight has been used widely [68]. This technique resolves the window size problem by creating a weighting matrix for the local area before the matching step, which is similar to using a highly adaptive window.

Contributions of each pixel in the local window are weighted by the similarity and proximity to the center pixel of the window.

2.2.2.2 Phase based methods

Local phase information is a robust source for binocular disparity [69]. The input image can be considered as 2D signals and the objective of disparity estimation is to shift the 2D images so that they match in the horizontal orientation. The spatial shift in image domain can be detected as phase shift in frequency domain. The computation of binocular disparity can be formulated as phase matching by determining the phase shift between the left and right signals [70]. Local band pass filters such as Gabor filters were widely used to reconstruct the local phases [70-71]. A one dimensional complex Gabor filter is shown in Equation (10).

$$\begin{aligned} G(x - x_0) &= e^{-\frac{(x-x_0)^2}{2\sigma^2}} e^{i\omega_0(x-x_0)} \\ &= e^{-\frac{(x-x_0)^2}{2\sigma^2}} \cos(w_0(x - x_0)) + ie^{-\frac{(x-x_0)^2}{2\sigma^2}} \sin(w_0(x - x_0)) \end{aligned} \quad (10)$$

The real and imaginary parts of the complex Gabor filter are also known as even and odd Gabor filters which have symmetric and asymmetric shape respectively. The response of filtering an image patch with Gabor filters is shown in Equation (11).

$$C(x_0, y_0) = \int I(x, y_0) G(x_0 - x) dx \quad (11)$$

The complex phase can be reconstructed as the phase angle in the complex plane, based on the real and imaginary component of the response.

$$\phi = \arg[C] = \arctan \left[\frac{\text{Im}(C)}{\text{Re}(C)} \right] \quad (12)$$

The complex phase difference between two image patches derived from the Gabor filter responses is divided by the instantaneous frequency to get the disparity.

$$D = \frac{\phi_L - \phi_R}{\omega_0} = \frac{\Delta\phi}{\omega_0} \quad (13)$$

Theimer et al developed a binocular vergence control model using complex Gabor responses [43, 72]. The minimization of global disparity of the full image was used for vergence control. If there are multiple depth regions in the image, the vergence control direct the cameras to an average depth. A common problem of phase based method is the wrap around of phase. Due to the periodicity of phase, the detection can only be limited in half the wavelength of the Gabor filters [71]. If a larger wavelength (spatial period) is used, the range of disparity is larger but the accuracy may suffer. For this reason, the coarse-to-fine processing methodology was often used to obtain both a larger disparity range and an improved speed [44, 56]. Hansen et al used a phase-based approach to compute multi-resolution disparity maps and the disparity at the center of image was used for vergence control [44].

2.2.2.3 Disparity energy model

There are many disparity estimation models based on biological and physiological studies of the primate vision system [45, 54, 73-74]. The disparity tuning responses of binocular cells have been investigated. Poggio and Fischer classified the binocular cells in the primary visual cortex into four groups: tuned excitatory (TE), tuned inhibitory (TI), near and far cells [75]. TE cells are tuned to zero disparity and excited by stimuli at the fixation distance. TI cells are the reverse of TE cells. Near cells are sensitive to stimuli nearer than the fixation distance and far cells are tuned to stimuli farther than the fixation distance. Later experiments showed a continuous tuned disparity range of the binocular neurons and disparity energy model was proposed to describe the binocular neurons [76-78]. The disparity energy model has thus far been considered the most neurophysiologically-faithful representation of the stereo mechanism in the mammalian

visual cortex. It was shown to be effective in estimating disparity in synthetic stereogram and real world images [54, 74, 79]. The disparity energy model describes the occurrences within the early visual cortex during binocular vision. It is likely that there are other mechanisms in the higher cortex that consolidate the responses and provide more robustness to stereo perception.

The disparity energy model consists of a population of disparity tuned complex cells that respond to a range of disparities. Each complex cell is constructed by a quadrature pair of binocular simple cells and their receptive profiles are modeled as 2D Gabor functions [80]. The binocular simple cells have been modeled differently where variations may exist either in the spatial location or phases in the left and right receptive fields. These models are respectively known as the position-shift and the phase-shift disparity energy models. The disparity of a certain location in the binocular visual input is then determined by the position or phase shift of the complex cell with strongest response. Figure 13 shows a complex cell. The position or phase difference between the Gabor filter G_L and G_R determines the preferred disparity of this disparity energy unit.

Figure 13 A disparity energy unit

While this method provides a faithful model for disparity estimation, the computational complexity of the model impedes the use of it in real time robotic systems unless it can be realized in parallel processing hardware. This has been done by neuromorphic implementations of the disparity energy model [49]. The disparity energy model and the direct phase reconstruction method share the same theoretical foundation since they are all based on Gabor filters to detect disparity from the band pass filtered responses. The difference between these two models lies in that the disparity energy model does not require explicit calculation of the phases. The disparity is however determined by the characteristics of the responses of a population of binocular neurons.

2.2.2.4 Cepstrum based methods

It is known that seated within the primary visual cortex, the visual information from the left eye and the right eye are sliced and interlaced, forming a structure called the ocular dominance columns. This led Yeshurun and Schwartz to develop the cepstrum based disparity estimation model [45]. This method interlaces the visual information from the two images, slice-by-slice in the horizontal direction, similar to the organization of ocular dominance columns in the visual cortex. The interlaced images are processed by the cepstral filter. The cepstral filtering response contains various correlation peaks and the disparities between the features in a binocular scene are estimated from the positions of the correlation peaks in the filtered image.

Cepstral filters have been used in auditory signal processing for detecting auditory echo [45, 81]. The cepstrum is the inverse Fourier transform of the log of the power spectrum. In the filtering process, an image column S in the left image is interlaced with an image column at the corresponding position of the right image. The width of the image

columns is D and the horizontal disparity between the two image columns is d . The input signal is just a window of width $2D$ containing the two image columns and can be expressed as Equation (14).

$$f(x, y) = S(x, y) * [\delta(x, y) + \delta(x - D - d, y)] \quad (14)$$

The logarithm of the Fourier transform is shown in Equation (15).

$$\log F(u, v) = \log S(u, v) + \log(1 + e^{-i\pi(D+d)u}) \quad (15)$$

The power spectrum of the second item can be expressed as Equation (16).

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{\delta(x - n(D + d))}{n} \quad (16)$$

This corresponds to various correlation peaks at multiples of $D+d$. The disparity d is expected to be less than half of the width of the image columns. Therefore the disparity can be determined by the peak response in the interval $[D/2, 3D/2]$ in the cepstrum plane. Furthermore, if objects exist at different depths in the input, the cepstrum will reflect multiple correlation peaks where each peak position corresponds to a disparity between a different set of features in the input. It was shown that cepstral filtering performs well in the presence of noise and other image degradations [45]. Cepstral filtering has been used in many vergence control systems [38, 82]. It was shown to be closely related to the phase correlation methods as the cepstrum can be thought of as autocorrelation with an adaptive nonlinear prefilter [57].

2.2.2.5 Log polar transformation based methods

The visual-cortical mapping in the primary visual cortex has been found to exhibit a log polar transformation structure [14-18]. Comparing with the Cartesian images, log polar images magnify the image center and diminish the periphery of the original image. This gives us a natural possibility of estimating the disparity and controlling vergence in

cluttered environments, where the foreground and background may have a wide gap of disparities. The spatial variant transformation is expected to show advantages over Cartesian images for the vergence control task, which requires a precise matching in the fovea. Log polar transformation has been used successfully for disparity estimation and vergence control [5, 46, 59, 83]. Sandini et al presented a binocular tracking system with vergence capabilities using log polar camera sensors [5, 84]. Divergence motion estimation and derivative correlation index in log polar images were used to control vergence. Zero disparity filtering in log polar domain was used to segment the foveated target. Correlation based vergence control models were also developed using log polar images transformed from normal Cartesian images [46, 59, 85].

2.2.2.6 Coarse to fine processing

Despite the performance of each correlation measures, a major issue of searching based disparity estimation methods is the limited working range. When the disparity existing in the image exceeds the working range, the algorithms are not able to get a reliable disparity. This occurs often when the system is altering the fixation point from a near proximity to a distant position. One approach to solve this is to use multi resolution image pyramid and process from coarser to finer levels [37-38]. The coarser scales allow the systems to handle large disparities effectively by limiting the number of possible matches while the finer scales systematically provide higher resolution [38]. The system presented by Taylor et al [38] is an example of cepstral filtering based hierarchical architecture where coarse-to-fine processing is realized through a foveated pyramid of the original image. The working range in this case is enlarged but still limited by the pyramid levels and search range at each level. Although exhaustive search can be used to make the working range equivalent to the view range, it is not computationally

efficient and noise may also be introduced into the estimation due to excessive uncorrelated visual information. Since spatial variant transform such as the log polar transform can give a stable matching on fovea, a plausible method of attaining both a wider working range and stable vergence on fovea is to use spatial variant transforms together with coarse-to-fine processing.

2.2.2.7 Version and vergence

Version and vergence control are the conjugate and disjunctive eye movements respectively [86]. Version movement is the binocular eye movements in the same direction while vergence movement is the eye movements in opposite directions through equal angles [87]. Vergence is typically bounded by a convergence of the line of sights at the front of the eyes (fixated) and a maximum divergence where both line-of-sights are parallel (looking towards infinity). Hering's law of equal innervation proposes that conjugacy of saccades is due to innate connections in which the eye muscles responsible for each eye's movements are innervated equally [88]. This means the movements of the two eyes are identical for a certain vergence or version control signal. This law presumes complete independence of the conjugate and disjunctive eye movements. It also states that apparent monocular eye movements are actually the mathematical summation of conjugate version and disjunctive vergence eye movements [88]. Natural binocular vision in an environment with objects at different depths and orientations involves both version and vergence control [23]. Yarbus studied the composite eye movements under this situation. Saccades were shown as having been additively superimposed upon a slower ongoing vergence movement [23]. However, there has been debate over the correctness of superimposition in vergence and version movements by subsequent research. It was shown that the binocular vergence system

seems to remain dormant during the major saccade [89]. Later experiments showed that the location of the new target, as seen by one eye, will lead to the generation of a saccade for that same eye, so as to produce good alignment with the new target [90]. The other eye however is left to achieve its foveation at a considerably later time. Hence, the saccades of the better-aligned eye seem often to have been programmed predominantly by monocular considerations [90]. This pre-dominant saccade and post-saccade vergence scheme is the basis for this thesis's investigations which eventually provides the mechanism of designing the binocular vision system.

Figure 14 The saccade-vergence cycle: (a) monocular detection of target, (b) saccade through parallel version movement and (c) vergence movement

As shown in Figure 14, this is evident in the implementation, where the version angle of an artificial system is first determined monocularly by one camera, functionally referenced as the 'master' camera. This monocular determination of the view is used as an engineering solution of obtaining a global view. The version angle then effects the vergence performed as a 'slave' operation in the other camera [87]. In this case, fixation consists of two separate movements with a decoupled control [25]. Many existing vergence control systems have implemented such a master-slave configuration [37, 39-40, 82].

2.3 The FLANN and KFLANN neural network

Two neural networks, FLANN and KFLANN are introduced here. The FLANN network was used in the CogV system in an image segmentation model for saccade generation. The KFLANN network was used in a Gabor signature based object recognition model for classification of objects.

2.3.1 The Fast Learning Artificial Neural Network

The Fast Learning Artificial Neural Network (FLANN) is an algorithm developed to perform fast learning of patterns. The FLANN network has an architecture similar to the Adaptive Resonance Theory (ART1) network [91]. It continuously creates new clusters for data not falling into any existing clusters. It achieves learning in a single epoch and produces consistently a good classification behavior [92]. This makes it suitable for real time applications such as image processing in robot vision.

The clustering process of FLANN is controlled by the tolerance δ_d for each dimension d and the vigilance threshold ρ . The tolerance setting defines the size of the generated clusters and is normally set according to the standard deviations of the training data for each dimension [92]. For an input data point and each existing cluster, a vigilance test is performed to determine whether to create a new cluster for this data or assign it to an existing cluster. A dimension of this data point and a cluster centroid is said to be matching if their difference in this dimension is less than the tolerance. If the ratio of the number of matching dimension over the number of total dimension is above the vigilance threshold, this cluster is a matching cluster of this data point. As there may be more than one matching clusters for an input data, the input data will be assigned to the nearest matching cluster with minimum Euclidean distance. If there is no matching

cluster, a new cluster will be created with this data as centroid. Note that the centroid of each cluster is actually an input data pattern. The FLANN algorithm is shown below [93].

The FLANN Algorithm

Step 1 Network Initialization: Given a dataset of n samples of each d features, set the preferred vigilance value ρ (recommended between 0.8 and 1.0). Set data index $k=1$. Calculate σ_d for all features in the dataset and assign tolerances $\delta_d = \sigma_d$.

$$\delta_d = \sigma_d = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_{kd} - \mu_d)^2}$$

Step 2 If ($k \neq n$) Present the k_{th} pattern x_k to the input layer. Else **Terminate**.

If (the network does not have any output node) GOTO **Step 6**.

Step 3 Determine all possible matching output nodes using

$$\frac{\sum_{i=1}^d D \left[\delta_i^2 - (w_{ji} - x_{ki})^2 \right]}{d} \geq \rho \text{ where } D[a] = \begin{cases} 1, & \text{if } a > 0 \\ 0, & \text{if } a \leq 0 \end{cases}$$

D is the discriminant function imposed on the difference between the tolerance and the distance between the node weight and the data in each dimension.

If (there is no matching node) GOTO **Step 6**.

Step 4 Maintain a list of all matching nodes that fulfilled **Step 3**. Using the winner-take-all criteria, determine a single winner node from the matching list.

$$winner = \min_j \left[\sum_{i=1}^d (w_{ji} - x_{ki})^2 \right]$$

Step 5 Assign the current pattern x_k to the winner node. GOTO **Step 2**.

Step 6 Create new output node j . Assign the node weight $w_j = x_k$. Assign x_k to node j and GOTO **Step 2**.

Notations: ρ is the vigilance value, δ_d is the tolerance for the d_{th} feature of the input space and w_j is the stored node weight vector.

2.3.2 The K-iterations Fast Learning Artificial Neural Network

The K-iterations Fast Learning Artificial Neural Network (KFLANN) is an iterative version of the FLANN network. Centroid stability is achieved through reshuffling the centroids to the top of the dataset after each iteration of learning. In the next iteration, the centroids will come into the network first and creates initial clusters. The centroids stabilize fast, normally in less than 5 iterations. The KFLANN algorithm is shown below [93].

The KFLANN Algorithm

Step 1 Network Initialization: Given a dataset of n samples of each d features, set the preferred vigilance value ρ (recommended between 0.8 and 1.0). Set iteration count = 1 and set data index $k=1$. Calculate σ_d for all features in the dataset and assign tolerances $\delta_d = \sigma_d$.

$$\delta_d = \sigma_d = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_{kd} - \mu_d)^2}$$

Step 2 If ($k \leq n$) Present the k_{th} pattern x_k to the input layer.

If (the network does not have any output node) GOTO **Step 8**.

Step 3 Determine all possible matching output nodes using

$$\frac{\sum_{i=1}^d D \left[\delta_i^2 - (w_{ji} - x_{ki})^2 \right]}{d} \geq \rho \text{ where } D[a] = \begin{cases} 1, & \text{if } a > 0 \\ 0, & \text{if } a \leq 0 \end{cases}$$

D is the discriminant function imposed on the difference between the tolerance and the distance between the node weight and the data in each dimension.

If (there is no matching node) GOTO **Step 8**.

Step 4 Maintain a list of all matching nodes that fulfilled **Step 3**. Using the winner take all criteria, determine a single winner node from the matching list.

$$winner = \min_j \left[\sum_{i=1}^d (w_{ji} - x_{ki})^2 \right]$$

Step 5 Assign the current pattern x_k to the winner node.

Step 6 If ($k = n$) Compute means for each cluster and identify a single cluster point closest to the mean as the cluster centroid.

If (collection of centroids are consistent with previous iteration or iteration count > 5) **Terminate**.

Else, increase iteration count by 1 and GOTO **Step 7**.

Else

Increase k by 1. GOTO **Step 2**.

Step 7 Reshuffle all centroids to the top of the dataset. Set $k=1$. Erase all output nodes.

GOTO **Step 2**.

Step 8 Create new output node j . Assign the node weight $w_j=x_k$, Assign x_k to node j and

GOTO **Step 2**.

Notations: ρ is the Vigilance value, δ_d is the tolerance for the d_{th} feature of the input space and w_j is the stored node weight vector.

3 The CogV System

A binocular vision system, CogV (Cognitive Vision), for simulating the human eye movements was designed and developed to provide the functional eye saccade and vergence for a pair of cameras. The developed saccade generation and vergence control models cooperate to provide the systematic means for the autonomous exploration of the scene. All investigations leverage on the hypothesis that the bio-inspired binocular visual perception can provide the necessary clues to develop a more robust perception system for scene understanding.

In this chapter, Section 3.1 gives a general introduction to the hardware platform and describes the models developed for the CogV system. Section 3.2 presents the developed saccade generation models. A watershed segmentation model and a FLANN based segmentation model were developed for saccade generation. Section 3.3 presents an object recognition model by Gabor filters. This model was developed during the investigation into the Gabor filters which describes the receptive fields of simple cells in the primary visual cortex. The saccade generation and object recognition models are supportive models for the investigations into vergence control. Therefore they are just generally introduced in this chapter.

3.1 An overview of the CogV system

Figure 15 The CogV system

The CogV system consists of a pair of cameras, each mounted on independent pan-tilt controls with 2 degrees of freedom, shown in Figure 15. The cameras are JAI CV-M7+ progressive scan cameras with 2/3" CCD image sensor and an 8mm C-mount (back focal distance of 17.53mm) lens. These capture images of maximum resolution 1380×1030 corresponding to horizontal angle of view of 64°. The cameras have no auto-focus and zooming. The pan-tilt control units are Directed Perception PTU-D46-17 Pan-Tilt control units. The motors have the capability to rotate at a speed of 300°/second with a resolution of 3.086 arc minute (0.0514°). For the purpose of all experiments conducted in this thesis, the baseline distance between the two cameras of the platform was set to 24cm. The cameras are connected to two Coreco PC2-Camlink frame grabbers. The pan-tilt control interface to the computer is RS232 serial port. The developed applications run on a desktop computer with Intel Xeon 3.2 GHz CPU and 2G RAM.

The system is designed to retrieve regions of interest in the scene, initiate saccades onto selected positions, and carry out vergence control on these positions. Figure 16 shows the flowchart of the CogV system. Several models in the system cooperate with each other to form a complete system simulating saccade and vergence. In this system, the SC (superior colliculus) model is responsible for generating interesting positions from the

image, maintaining a global map of the positions and providing the fixation positions. The VC (visual cortex) model provides vergence capability and let the two cameras verge on the same point in space.

Figure 16 The flowchart of the CogV system

The system assigns one camera as master camera and the other as slave. From Figure 16, the SC model provides saccadic landing spots from the image captured by the master camera. A spatial map of the environment is constructed from the saccade generation model. During each saccade, a position in this SC map is selected as a target position. The parallel motor control model performs an initial major saccade, where the two cameras move rapidly in the same direction to the proximity of the attended region. After the initial saccadic burst of motion, the master camera fixates accurately on the object while the slave camera is directed only to the proximity of the master camera's fixation point. At this instance, vergence may yet be achieved.

The system subsequently activates a secondary process within the VC model for a precise vergence control. The VC model estimates the disparity of the image center and iteratively corrects the slave camera so that the area of interest appears at the fovea of both the master and the slave cameras. Once verged, the position and features of object

within the fovea can be registered by the system through a training process or a heuristically automated process to mimic child-like visual-babble-learning.

An inhibition-of-return (IOR) is implemented for cases when CogV is not expected to return to the same position during the saccade phase. At the new fixation position, saccade generation is carried out again and any new saccadic positions different from those stored in the SC map will be put into the SC map. The system then proceeds to generate the next saccade from the SC map and fixates the cameras on a new location of interest. Privilege is given to the nearest saccade candidate position in the SC map. The master-slave configuration of binocular cameras of the CogV is not fully equivalent to the human vision system but it is effective and efficient for design and implementation of the saccade and vergence for artificial vision systems. Through a repeated combination of saccade and vergence, the system is able to perform an autonomous exploration of the environment.

This thesis focuses on modeling these saccadic-vergence processes and explores the efficiency of each model. Figure 17 charts the work done in this thesis and a deeper emphasis on the development of vergence control models for the CogV system is evident. Saccade generation models were developed using watershed and FLANN segmentation techniques. A Gabor signature based object recognition model was also developed. Both the saccade and recognition models are supporting models of the CogV system. The developed saccade and recognition models are presented in this chapter. The vergence control models are presented as separate chapters after this chapter.

Figure 17 Developed models for the CogV system

3.2 Saccade

From studies in neuroscience, the superior colliculus (SC) is attributed for its heavy involvement in saccadic eye movements [9, 94]. The colliculus is thought to coordinate the head and eye towards something seen or heard. In vision based saccade generation, visual processing models that detect interesting positions or regions in the image could be used for saccade generation. The detection can be based on attributes such as color, edge, texture, motion, shape or combination of features. Computational visual attention models have also been developed in recent years, such as [28].

This section presents the developed saccade generation models in the CogV system. The first model presented is a watershed model for image segmentation. This model utilizes the edge features in the image and simulates the object detection as a water filling process. After this a Fast Learning Artificial Neural Network (FLANN) based image segmentation model is presented. The segment centers were passed through a selection

process to retrieve informative segments. Two techniques were used for selection of segments. The first is through checking local information around the segment centers to remove irrelevant positions with insufficient variations or insufficient amount of features. The second is through checking the color attention associated with the segment centers. In the system, the order of saccade is governed by the attention level and the proximity to the current fixation position or the image center. The saccadic position which passes through the attention threshold and is nearest to the current fixation position will be generated as the next saccade. The saccade generation process is a process to select important positions in the image while the relative importance of the selected positions is not discussed in the current scope of work.

3.2.1 The watershed model for saccade generation

This section presents a simplified model based on a watershed segmentation algorithm to emulate the superior colliculus (SC). The watershed model gets its inspiration from water collecting in low lying areas surrounded by high terrain during a rainfall [95]. If water is poured into an enclosed space then it would accumulate and the water level will rise. The assumption is that objects inside the scene have enclosed boundaries and there are no objects of a very large size.

The proposed watershed model has two layers. The first layer is the ganglion layer which consists of ganglion cells performing contrast detection operation. The ganglion layer has the same size with the input image. The output of ganglion layer is an edge map from the input image. The second layer is processing layer. Edge points from the ganglion layer act as terrain cells which blocks water signal flowing. The non-edge points act as water cells which simulate flow of water and subject to constant increase of water

signal strength. The terrain cells and water cells all correspond to pixel locations in the original image. The second layer forms a 2D grid to control the flow of the water signal. As shown in Figure 18, a water cell receives signal from its 8-neighbors. Water flows to the cell if there is a path from this cell to one of its 8-neighbors which is a water cell.

Figure 18 Water flow in watershed model

In the second layer, excitatory inputs are constantly fed into the water cells to simulate water fall. At the boundaries of this water-flow cell matrix (boundaries of image) are sink cells that act as sink holes, removing the signal strength from the whole matrix. The signal strength or water level in the water cells of the processing layer is an indication towards the cell being a part of an object or not.

In the implementation, water flowing is simulated by an averaging operation of each water cell and its water cell neighbors. For example, the excitatory input for a water cell W in the second layer is A_w . W has K water cell neighbors which can give water flow to W , each with a water level of A_k ($k=1,..,K$). Then the new water level of W is set to $(A_w + A_1 + ... + A_K) / (K + 1)$. The updating of water levels are systematically carried out from left to right, right to left, top to bottom and bottom to top of the processing layer. A saturation level is set for the water level to avoid increasing to infinity. After a certain number of iterations, some cells will gain at a high water level and some will not.

An additional step is added to the model to retrieve the centers of the connected regions. This is done through thresholding the water strength to determine those active water cells with high water level and systematically traversing the image and assigning neighboring active water cells the same index. The indices indicate the objects in the scene. The positions of each indexed region can be easily obtained as the mean coordinates of the water cells in the region. These positions are used as saccade generation source. Figure 19 shows a simple scenario of how the watershed model works. Figure 19c shows the water strength in the processing layer and Figure 19d shows the detected object positions as bright dots.

Figure 19 Watershed on a real scene: (a) input image, (b) edges, (c) water strength and (d) object positions

The model is able to locate positively present objects and is also resilient to noisy signals arriving from the ganglion cells. As shown in Figure 20a, in noisy environments where edges of objects are indistinct, ganglion cells at the edges may not activate. This causes a break in the edge-link, providing inner cells with additional paths to the sink cells. However, if the rate of signal strength introduction is higher than the rate of signal decay, the accumulated signal strength will increase, showing the presence of an object. The system is also resilient to false edges caused by noise as these will not connect to form any closed sections. The simulation in Figure 20 illustrates the model's resistance to noise.

Figure 20 Resistance to noise of the watershed model: (a) edges, (b) water strength and (c) cells crossing threshold show presence of objects

While the watershed model is fairly robust, the assumption that objects present a near close-loop object contour characteristic may not always hold true. In the event a large object or long edges form a closed-loop at the background, the foreground objects will entirely be merged into a single water pool. The developed saccade generation model usually performs well in an experimental environment where the background is smooth and foreground objects are not too large with small close edge loops. To overcome this limitation, a new model was developed based on color image segmentation.

3.2.2 The FLANN segmentation model for saccade generation

This section introduces another basic functional representation of the superior colliculus module that was implemented using a Fast Learning Artificial Neural Network (FLANN) image segmentation method [96]. Figure 21 shows the SC model. The input image is segmented by the FLANN segmentation model and then passed through a segment selection process to determine which segment centers are suitable for saccade generation.

Figure 21 The SC model for saccade generation

3.2.2.1 FLANN Segmentation

The process of FLANN segmentation is presented in Figure 22. The input image is initially preprocessed using an average operator to smooth the image and image pixels were then transformed into the RGBSV color space. After this, the image data is passed to a FLANN network for clustering. According to the generated class labels, the image pixels are assigned initial segment numbers. Color regions in the image may be split to small segments due to the input data sequence. The initial segments generated are merged according to their cluster label and spatial adjacency. The merged segments are the final segmentation result.

Figure 22 The FLANN segmentation model

The *RGB* (red, green and blue) color model uses the three primary colors of light to describe a color. Each is represented by a gray level proportional to the component's intensity. Another well received color space model is the *HSV* (hue, saturation and value) model which provides more accurate description of perceptual color representation [97]. Hue represents the fundamental or dominant color as perceived by an observer. Saturation refers to relative purity or the amount of white light mixed with a hue. Value stores the luminance intensity information. In the algorithm proposed, the clustering of colors is in the *RGBSV* space, which is a combination of *RGB* and *HSV* spaces. Hue is removed from consideration due to its large variation at low *RGB* values. The full *RGB* configuration is kept and additional *S* and *V* information is included to assist in the

FLANN vigilance testing. It is believed that the redundant information presented in the new *RGBSV* space provides room for better segmentation disparity. The conversion from *RGB* to *SV* components is shown in Equation (17).

$$\begin{aligned}
MAX &= \max(R, G, B) \\
MIN &= \min(R, G, B) \\
S &= \begin{cases} 0, & \text{if } MAX = 0 \\ 1 - \frac{MIN}{MAX}, & \text{if } MAX \neq 0 \end{cases} \\
V &= MAX
\end{aligned} \tag{17}$$

The *RGBSV* data are normalized to the range [0,255]. The Vigilance of FLANN is set to 0.8, which means at least 4 out of 5 dimensions should be within the tolerance range so that a pixel can be assigned to a cluster. This provides a relaxation of the cluster space so that noise in the data can be compensated. The tolerances are set empirically to 60 for all 5 dimensions. Hence the radius of the clusters can reach 60 for each dimension. The entire color space will be divided into a group of subspaces, which can be considered as a color reduction process.

Through the FLANN clustering, the image can be partitioned into distinct regions. Regions are connected pixels bearing the same cluster index. The process of segmentation takes the data clustered by the FLANN and fuses pixels within a color cluster that are spatially adjacent into the same image segment. The adjacency is determined by defining a neighborhood area of each pixel. If a pixel *B* falls in the neighborhood of a pixel *A* and has the same cluster label with *A*, the pixel *B* is heuristically assigned to belong to the same image segment with *A*. The CogV system uses an input image size of 200x200 and a segment neighborhood area of 21x21 pixels. This segmentation assigning process is shown in the pseudo code below.

Define: $C(x)$ is the cluster label of pixel x
 $S(x)$ is the segment label of pixel x
 $N(x)$ is the neighbourhood area of pixel x
For each pixel x
 If $S(x) == NULL$
 Assign a new segment label to x
 End If
 For each $x_n \in N(x)$
 If $C(x_n) == C(x)$ and $S(x_n) == NULL$
 $S(x_n) = S(x)$
 End If
 End For
End For

Since the input sequence of image data is sequential, systematically evaluating from left to right and top to bottom, it is inevitable that a single concave region may be fragmented into disjoint segments. The pseudo code below is a post segmentation merging process necessary to resolve this fragmentation.

Define: $N_8(x)$ is the 8 – neighbourhood of x
For each pixel x
 For each $x_n \in N_8(x)$
 If $C(x_n) == C(x)$ and $S(x_n) \neq S(x)$
 Merge the segments which x_n and x belong to
 End If
 End For
End For

After this merging process, the center of segment is calculated from an average of all pixels in the segment. Adjacent small segments generated are further merged to form bigger segments. The criteria for merging are similar to that practiced in the segmentation of pixels. Small segments whose centers are near to each other are merged to form bigger segments. The neighborhood area is again defined as 21x21. The

centers of these bigger segments are then marked as possible candidates for saccade generation. This approach is similar to that presented in [98]. For a concave segment, the possibility of having the center outside the segment body arises. To resolve this problem, segments that have centers within a given proximity are merged. Assuming a simple background, the merged segment centers provide promising representatives of objects in the environment.

For an input image of n pixels which is clustered to c clusters, the computational complexity of FLANN clustering is $O(nc)$ because it is a one-epoch algorithm. The number of clusters is normally less than 50 for our current setting of vigilance and tolerance. If the neighborhood size is $h \times h$ and the number of pairs of segments merged is p , the segmentation and merging costs $O(n(h^2+p))$. p is in the order of tens because most regions in an image will not be split by the segmentation algorithm. Comparing to iterative clustering methods such as k-means, the FLANN network provides an efficient way to profile colors in the image.

3.2.2.2 Selection of segments by checking local variation

This section presents a segment selection approach by checking local information. The saccade candidate positions generated by FLANN segmentation are assessed to satisfy properties of a potential fixation point. The first criterion is the presence of a significant level of intensity variations and edge information in a local neighborhood. The intensity variation is determined by the entropy of the intensities in the local neighborhood. The entropy of an image region indicates the randomness or variations in the region. It is assumed that potential object positions have large variations to distinguish the object from background. The entropy for a region of the image is produced from the intensity

histogram of this region, defined in Equation (18). In the CogV system, the neighborhood is set to 21x21 pixels. The threshold for entropy was empirically set to 2.0.

$$S = -\sum_i P(i) \log(P(i)) \quad (18)$$

where $P(i)$ is the probability of intensity i in the region

After checking entropy, edge information from the image is used to suppress non-informative regions. An odd-type Gabor filter bank is used to filter and construct an edge image. Gabor filters are used primarily for their succinct representation that can mimic profiles of the simple cells in the human visual system and furthermore, they are tolerant to noise. Equation (19) shows the equation for 2D Gabor filters. The filter bank used here is a filter bank of 8-orientations (θ evenly distributed from 0 to π), 16x16 size, $\lambda=4$, $\sigma_x=1.25$, $\sigma_y=1.25$ and $\varphi=\pi/2$. The edge strength of a point is the maximum Gabor filter response among the 8 orientations for this point. The edge detection threshold is set to 0.2 of the maximum response in the filtered edge image. If an area is marred by insufficient edge information at the segment center, the position will be rejected to be a point of interest. For the local neighborhood of 41x41 pixels, we expect the position to have at least 20 edge points (radius of the neighborhood) to be a saccade position.

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \cos\left(\frac{2\pi x'}{\lambda} + \varphi\right) \quad (19)$$

$$x' = (x - x_0) \cos(\theta) - (y - y_0) \sin(\theta)$$

$$y' = (x - x_0) \sin(\theta) + (y - y_0) \cos(\theta)$$

Apart from seeking or avoiding uninteresting regions, a spatial pan-tilt boundary constraint for the saccade was implemented. Object positions lying outside the boundary will not be considered, causing the cameras to ignore those saccade positions. This boundary possesses the maximum horizontal angle of view of 30° from the vertical meridian and a vertical angle of view of 15°. This boundary avoids the distortion of

images due to the static baseline. This is similar to the biological system where viewing beyond a certain angle will normally be accompanied by head movements instead of just eye movements [9]. After fulfilling all constraints, a set of informative positions are generated and consolidated as saccade target locations.

Figure 23 shows the results of the SC model on a real scene. In the image, several objects at different depths were placed in front of the system. The SC model retrieved regions and enclosed selected regions in rectangular blocks. Regions that contain less variation could be suppressed by raising the entropy based threshold (Figure 23b). The spatial constraints removed positions that were out of range (Figure 23c). Regions containing fewer edges were removed by checking the Gabor edge information (Figure 23d). A list of informative positions suitable for saccade generation was eventually generated (Figure 23d). On the current platform (Intel CPU of 3.2GHz and 2G RAM), this FLANN segmentation model processes the 200×200 image in about 0.8s and the local information check takes about 1.5s given about 10 regions to check.

Figure 23 FLANN segmentation of scene images: (a) original image, (b) FLANN segmentation with entropy threshold, (c) add tilt constraints and (d) add Gabor edge threshold

3.2.2.3 Selection of segments by color attention

To realize a similar functionality of the local information checking, a color attention aided saccade retrieval method is presented here. Combining the segmentation results

and a color attention model, informative segments in the scene can be selected. The color attention model comes from Itti and Koch's visual attention model [28]. Orientation saliency is also an important component in this model. However, as the segmentation model segments the image according to color, the orientation component is omitted and this also reduces the computational load significantly.

In our implementation, the input image has a resolution of 200×200 . Each pyramid consists of 6 levels with scaling ratio of 0.5 between successive levels (level 1 being the finest and 6 the coarsest). The center scale c is in the range $[1, 2]$ and the surround scale s is in $[c+3, c+4]$. The center-surround maps were resized to the level 3 resolution (50×50) and summed to produce the individual feature saliency map. The feature saliency maps were then summed to produce the final saliency map. The final attention map describes the conspicuity or saliency in the image and was used to reject segments with low saliency, leaving only segments with high attention for saccade generation. In this way, the large smooth regions with less variation are eliminated, providing a similar function with the local information checking using entropy and Gabor filtering. Figure 24 shows the saccade generation process. The original image was segmented and the attention levels of segment centers were used to determine whether it is a position worth to move eyes on. Figure 24d shows the final result. The attention threshold is set to 0.4 in the final map normalized to the range $[0, 1]$. On the current platform, the segmentation and attention based saccade generation takes about 850ms (the attention map construction costs only 50ms), which is much faster than the approach in section 3.2.2.2.

Figure 24 Saccade generation by segmentation and attention: (a) original image, (b) FLANN segmentation, (c) attention map and (d) candidate saccades

A saccade experiment was conducted using the master camera of the CogV system. The sequence of saccade generated is based on the principle that privilege is given to objects nearer to the current fixation point. Inhibition of return (modeled as an inhibition region) is included to prevent it from looking back. The motor positions were recorded down and transformed to the image domain and superimposed on the image captured at the beginning of the experiment. The saccade generated in a simple environment is shown in Figure 25a. The system started from the magnetic shapes on the wall and sequentially attended to the five objects on the table. Figure 25b shows another experiment in a lab scene. Ten saccades were generated and these covered a number of positions corresponding to several objects in the image, such as the lamp and the paper on the door. When the attention threshold is set to a lower value, more segments were selected and this is shown in Figure 25c. More objects were included in the saccade path. Some positions are out of the view of the image and are not shown in the figure.

Figure 25 Saccade generated with attention threshold $T_a=0.4$, $T_b = 0.4$ and $T_c = 0.2$

To compare the developed model with human saccade, Figure 26 shows results of a simulated saccade experiment from three human subjects. The picture of the scene was shown to the human subjects on a 17 inch computer monitor 50cm away. The human subjects were asked to use mouse to click the sequence of the landing points of their eyes on the image. A number of mouse clicks were collected and displayed on the image after the experiments. Although this is a very subjective experiment, it is still interesting to see that the saccade generated by the image segmentation model is partially compatible with those from the human subjects. The human subjects also tend to look for positions near the current fixation point. This agrees with our assumption that the selection of the next saccade prefers nearer positions.

Figure 26 Simulated saccade by accepting mouse clicks from human subjects

Through the investigations using the CogV system, it was discovered that the watershed model did not perform well in a cluttered environment. As such, the FLANN segmentation with visual attention based saccade generation is recommended for the saccade generation process.

3.2.3 Saccade integration

The human scans the visual world through fast saccadic eye movements. Transsaccadic integration is the process of integrating the visual information from the resultant sequences of images [99]. In the CogV system, the SC model maintains a global map to store the saccade positions generated. This global map is constructed incrementally, similar to the transsaccadic integration process where the human brain uses eye-position signals in interpreting retinal information [99]. Each time the system fixates on a new position, it generates interesting positions from the captured image and creates a local motor map with respect to the current fixation position. The local motor map can be integrated to the global map after offsetting by the current fixation position vector. The selection of saccade is always egocentric.

Figure 27 The locally constructed global motor map

As shown in Figure 27, at each fixation position, object positions in the current view are generated and stored in the global map. The system uses the fovea of the image as

reference and selects a position to fixate. The position offset in image plane provides the motor step between the current position and the expected target position. Errors due to perspective distortion and motor rotations can be limited by choosing saccade nearest to the current position, thus reducing the turning angle of the two eyes. This is similar to the human vision system where objects located more than 15° from the fixation point will invoke a head turning reflex and subjects will rotate their head to reduce the angle of the fovea to the new target fixation point [9].

The mapping between visual information and motor control is as follows. The lens and CCD is assumed to be a perspective camera model and the visual information falls on the image plane according to uniform angular mapping. Let the horizontal angle of view be A , the horizontal resolution of the captured image be M . Then each pixel corresponds to an angle of A/M° . Let the step resolution of the pan-tilt control units be S degrees/step, the mapping from pixel to pan-tilt control is $P=A/M/S$ pixels/step. In the CogV system $P=0.8927$ pixel/step. The visual distance in the image is used to derive the motor control signal when the camera is expected to move from one position to another position in the image.

3.3 Neural classification of objects based on Gabor signature

One objective of the CogV system is the recognition of objects under fixation. This section presents an object recognition model developed during the investigation of the CogV system. This model uses a combination of K-iterations Fast Learning Artificial Neural Network (KFLANN) and Gabor filters to create a Gabor signature classifier.

Simple cells in the visual cortex have spatially localized receptive fields, which consist of distinct elongated excitatory and inhibitory zones [13, 100]. Gabor filters have been considered the best profile for describing the receptive field of simple cells in the primary visual cortex [80]. It was shown that the receptive-field properties of mammalian cortical cells are well suited for representing the information contained in natural images [101]. Gabor filters, being sensitive to edges and lines, have been widely used in edge detection [102], corner detection [103-104] and contour detection [105]. Much of the current research have been focusing on the utilization of the Gabor filter to generate useful features for recognition or representation of the content of images [106-107]. Studies on Gabor filter design have also been used to produce Gabor filter banks that preserves maximum image information with minimal redundancy [108-109]. Due to the Gabor filter's ability to reduce dimensionality and yet maintain principal feature constructs, Gabor filters have been utilized in many recognition applications such as face recognition, image retrieval, fingerprint recognition, script identification and texture recognition or classification [108, 110-118]. Two elementary Gabor filters, the antisymmetric and symmetric Gabor filters are illustrated in Figure 28. They are also known as odd and even type Gabor filters.

Figure 28 Antisymmetric and symmetric Gabor filters

This section leverages on the excellent orientation abstraction properties of the Gabor filters to create a viable object signature that is invariant to scaling and rotation in the recognition of 2D planar objects.

3.3.1 Gabor filters

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \cos\left(\frac{2\pi x'}{\lambda} + \varphi\right) \quad (20)$$

$$x' = (x - x_0) \cos(\theta) - (y - y_0) \sin(\theta)$$

$$y' = (x - x_0) \sin(\theta) + (y - y_0) \cos(\theta)$$

Equation (20) shows the two-dimensional Gabor function. The pair (x_0, y_0) specifies the center of a receptive field in image coordinates. The standard deviation σ_x and σ_y of the Gaussian factor determines the size of the receptive field. The ratio $\gamma = \sigma_x/\sigma_y$ is the aspect ratio of the receptive field. σ_x can be determined by Equation (21) [115], where b is the spatial frequency bandwidth of the filter and λ is the wavelength.

$$\sigma_x = \frac{\lambda}{\pi} \sqrt{\frac{\ln 2}{2} \frac{2^b + 1}{2^b - 1}} \quad (21)$$

The eccentricity of the receptive field ellipse is determined by the aspect ratio γ . It has been found to vary in a limited range of $0.23 < \gamma < 0.92$ [119]. The ratio σ_x/λ determines the spatial frequency bandwidth b . The parameter θ ($[0, \pi)$) specifies the orientation of the filter, which is the counterclockwise rotation of the filter. θ is the orientation which is perpendicular to the parallel excitatory and inhibitory stripe zones of the filter [115]. The parameter φ ($(-\pi, \pi]$), which is a phase offset in the cosine term, determines the symmetry of the function g with respect to the center. For $\varphi=0$ and $\varphi=\pi$ the function is symmetric with respect to the center of the receptive field. For $\varphi=-\pi/2$ and $\varphi=\pi/2$, the function is antisymmetric. All other cases can be considered asymmetric mixtures of

these two [115]. If the spatial aspect ratio γ is fixed at 0.5, by varying the scale σ and orientation θ of the elementary Gabor filter, the filter banks in Figure 29 are obtained.

Figure 29 Gabor filter profiles with 4 scales and 8 orientations, normalized to $[0, 1]$: (a) antisymmetric Gabor filters and (b) symmetric Gabor filters

Figure 30 shows the results of Gabor edge detection applied to the Lena image [120]. The image is filtered by an antisymmetric Gabor filter bank with 12 orientations (θ evenly distributed in $[0, \pi)$). The maximum response of 12 orientations for each pixel is used to construct the edge image, shown in Equation (22).

$$E(x, y) = \max(|R(x, y, \theta)|) \quad (22)$$

where $R(x, y, \theta)$ is the θ orientation response at pixel (x, y)

The results show that Gabor filters are resilient to high frequency noise and gradual changes. For example, shadows affect little on the robustness of edge retrieval.

Figure 30 Edge detection by Gabor filters

Edge orientation detection using Gabor filters is by determining the strongest orientation response of each pixel in the image. Figure 31 shows the result of using both

types of Gabor filters for orientation detection. Antisymmetric and symmetric Gabor filter banks with 12 orientations (θ evenly distributed in $[0, \pi)$) were used. The following parameters were used: $b = 8$, $\gamma = 0.5$, $\lambda = 8$, and filter size was set at 32 pixels. The edge orientation θ at a pixel is determined by the strongest response from 12 orientations.

$$O(x, y) = \underset{\theta}{\operatorname{argmax}}(|R(x, y, \theta)|) \quad (23)$$

The edge orientation image is shown as grayscale image in Figure 31. Each orientation is represented by a unique gray level. In Figure 31b, the antisymmetric Gabor filters produces a strong and wide responses around the edge region. The orientation detection along the edge region is however thickened with overlaps with other orientations. In Figure 31c, symmetric Gabor filters produces sharp and accurate orientation detection results around the edge region with less overlaps and the final output is less noisy along the edge region. However, the thickened area has a wider width.

Figure 31 Edge orientation detection using Gabor filters: (a) input image, (b) orientation image by antisymmetric Gabor filtering and (c) orientation image by symmetric Gabor filtering

The orientations in the uniform regions of the image do not indicate any relevant edge orientation information because these regions gives almost equal responses for all orientations of Gabor filters. Therefore, the uniform regions with no edge information must be removed if edge orientation information is utilized to construct an object

signature. Similarly thickened regions with wrong edge orientation information along the edges should also be removed. These can be achieved by an edge region mask.

3.3.2 Edge region and orientation retrieval

In this section an edge information retrieval model is presented. Two problems were addressed in the model for the construction of an object signature. The first is how to retrieve accurate edge regions in the image. The second is how to efficiently determine the edge orientations in the edge region.

The proposed model consists of 3 Gabor filter banks. The parameter settings are shown in Table 2. This contains two antisymmetric Gabor filter banks and one symmetric Gabor filter bank. The LargeAsym and LargeSym filter banks are antisymmetric and symmetric filter banks with larger receptive fields (lower spatial frequencies). The SmallAsym filter bank has smaller receptive fields (higher spatial frequency). The difference between the receptive fields can be seen in Figure 32a, b.

Table 2 Parameter settings of Gabor filter banks

Parameters	SmallAsym	LargeAsym	LargeSym
b	4	8	8
γ	0.5	0.5	0.5
λ	4	8	8
φ	$\pi/2$	$\pi/2$	0
Size	20	20	20
σ_x	$\frac{\lambda}{\pi} \sqrt{\frac{\ln(2)}{2} \frac{2^b + 1}{2^b - 1}}$		
θ	12 orientations in $[0, \pi)$		

The LargeAsym Gabor filter bank has the specific function for edge detection while LargeSym Gabor filter bank determines edge orientations. The SmallAsym Gabor filter bank is used to produce a mask to extract the precise edge regions. The smaller bandwidth and wavelength produces a profile that can be thresholded to obtain a thin edge region near the edge peaks. This is illustrated in Figure 32d. With a threshold in the SmallAsym filters, it is possible to define an accurate edge region which bounds the edge orientation detection by the larger filters to the precise edge orientations in a thin strip along the edges.

Figure 32 (a) LargeAsym Gabor filter, (b) SmallAsym Gabor filter, (c) input image and (d) horizontal profiles of the responses of LargeAsym and SmallAsym Gabor filter banks

Figure 33 shows the flowchart illustrating the process of generating edge orientations. The left branch of the flowchart shows the construction of the edge region image from the SmallAsym Gabor filtering process. This is achieved by Equation (22). By defining a threshold, the smaller edge responses can be suppressed to obtain an edge region image. Here the Gabor edge strengths are normalized between $[0, 255]$ and according to

empirical experiments the threshold is set to 100. This edge region is used as a mask to retrieve the center regions along lines and edges in the image.

On the right branch of the flowchart, an antisymmetric Gabor filtering with a larger receptive field is used to obtain an initial edge image. This edge image is subsequently passed through symmetric Gabor filtering for the detection of edge orientations.

Figure 33 Edge orientation detection

In the orientation detection process, instead of using the strongest response for orientation detection, each orientation response is subtracted by the response in the perpendicular orientation. The orientation with the largest value after subtraction is the orientation of the edge. This is accomplished through Equation (24). This subtraction emphasized the oriented nature of the edges under the assumption that a strong oriented edge should have weaker responses in the perpendicular orientation.

$$O(x, y) = \begin{cases} \underset{\theta}{\operatorname{argmax}} \left(|R(x, y, \theta)| - \left| R\left(x, y, \theta + \frac{\pi}{2}\right) \right| \right), & \text{if } \theta \in \left[0, \frac{\pi}{2}\right) \\ \underset{\theta}{\operatorname{argmax}} \left(|R(x, y, \theta)| - \left| R\left(x, y, \theta - \frac{\pi}{2}\right) \right| \right), & \text{if } \theta \in \left[\frac{\pi}{2}, \pi\right) \end{cases} \quad (24)$$

The orientation detection result is illustrated in Figure 34a, in which each orientation is denoted as a unique gray level. The result in Figure 34a has less noise and less

orientation overlaps than those shown in Figure 31b,c. Using the previous constructed edge region image as a mask, it is possible to get the edge orientations for each edge pixel. Ignoring the border region, the final result after edge region masking, shown in Figure 34b, shows precise edge region and edge orientations. The edge orientations in the edge region are shown as arrows in Figure 34c.

Figure 34 (a) Edge orientation detection using antisymmetric followed by symmetric Gabor filter bank, (b) edge region retrieval and (c) edge orientations

3.3.3 Gabor signatures

A Gabor signature for 2D object recognition is generated based on the edge orientation information using a histogram based approach. The flowchart of the signature generation process is shown in Figure 35.

Figure 35 Gabor signature generation

With the help of a histogram, the orientation that is most prevalent in the object is selected as the principal orientation. The principal orientation is then used as an anchor reference that determines the signature's start position. All other orientations, regardless of their histogram strength of contribution will then be arranged in order of their angular distance from the principal orientation. Since the twelve orientations

always form a ring when arranged in their angular distances, the principal orientation can be perceived as a start point of the signature and the complete signature is formed by the subsequent neighboring orientations of the ring. The twelfth orientation will essentially be the neighboring point to the principal orientation on the other side of the starting direction. This shifting addresses the rotation invariance of recognizing objects in the 2D image. Scaling invariance can be obtained by storing the signatures as percentages.

Figure 36 Scaled and rotated images and their Gabor signatures: (a) original images and (b) images with Gaussian noise

Figure 36a shows the Gabor signatures from scaled and rotated images. The Gabor signatures were normalized by dividing each bin by the principle component. The signatures generated only shows slight difference. This strengthens the point that Gabor

signature reveals consistent information for object identification. Subsequently, Gaussian noise with a 0 mean and 0.01 standard deviation was introduced to the images. The images with Gaussian noise and the Gabor signatures are shown in Figure 36b. The Gabor signatures generated is very similar to those generated without Gaussian noise, illustrating the robustness to noise by Gabor filtering.

3.3.4 Classification of 2D objects using Gabor signatures

Figure 37 Sample objects

An object classification experiment was conducted using the Gabor signature. The test data was obtained from a series of images obtained from the five objects, shown in Figure 37. Each object was used to create images with 3 scaling factors (1, 1.25, and 1.5) and 10 rotational orientations ($[0, 0.9\pi]$ counter clock wise). A total of 30 images were generated for each object. The full image test set has 150 ($5 \times 3 \times 10$) images. Each was used to produce a 12-dimensional Gabor signature and a numeric label was used to tag the object. A second data set of 150 images was obtained by introducing Gaussian noise (0 mean and 0.01 standard deviation). This was used to evaluate if the Gabor signatures provide a reliable signature under noisy conditions. These two data sets were named **Dataset 1** (no noise) and **Dataset 2** (with Gaussian noise) respectively. The two data sets were fed into a clustering or classification algorithm of choice. This was also to evaluate the suitability of the clustering or classification algorithms. Two modes of testing were applied to the dataset. The first was a memory recall testing. The full test dataset was presented to the models for both training and testing. The second mode was a

generalization testing. The dataset was divided into a training set and a testing set. The training set was used by the models for training. Then the testing set was used to test the generalization capability of the models.

Four methods were used for the learning and classification of Gabor signatures to determine if the Gabor signatures provide useful information for distinguishing objects. The four methods are nearest neighbor classifier (NN), backpropagation (BP) neural network [121], K-Means [122] and the KFLANN network [93]. This also provided a means to compare the results obtained from the KFLANN and the other methods. The NN was implemented in Matlab by the author and the BP and K-Means models were implemented using the functions provided by the Neural Network Toolbox and Statistics Toolbox of Matlab. In this experiment, the tolerance values of KFLANN are set to the standard deviations of the training data. KFLANN are applied to the Gabor signature dataset using two vigilance (ρ) settings. In NN, the arithmetic mean of data samples belonging to each class was used as cluster centers. Testing data samples were classified to the 5 classes according to the nearest neighbor rule. In BP, the number of hidden neurons was set to 10, which was twice the expected number of clusters in the data. The training process is a supervised process. Both the data and the label were presented to the BP network for learning. The cluster label was considered as floating point data and testing was conducted by converting the output to the nearest integer for comparison with the labels. In K-Means, the parameter K is set to 5 and 10. When K=5, the cluster means of the 5 classes are used as the starting seeds. When K=10, each cluster is randomly divided into two halves. The mean of the 10 halves were used as starting seeds. The cluster label was determined by majority voting from the data in the cluster. Error was determined by those data that had different label from the cluster label.

Table 3 shows the memory recall testing results. KFLANN attained a clustering accuracy of 100% for both original image set and image set with Gaussian noise. The performance of K-Means is worse than the other methods. Overall speaking, the performance on the image set with no noise is slightly better. However, the performance on the image set with Gaussian noise is also robust. This again indicates the robustness of Gabor signatures against noise in the image.

Table 3 Clustering of Gabor signatures

Method	Settings	Dataset 1		Dataset 2	
		Number of Clusters	Accuracy	Number of Clusters	Accuracy
NN		5	93.33%	5	97.33%
BP	10 neurons	5	99.47%	5	94.67%
K-Means	K=5	5	92.67%	5	88.67%
K-Means	K=10	10	93.33%	10	98.67%
KFLANN	$\delta=\sigma, \rho=0.8$	12	96.67%	14	94.67%
KFLANN	$\delta=\sigma, \rho=0.9$	18	100.00%	22	100.00%

To further compare the performance of KFLANN for this signature classification task, the generalization capability of the KFLANN and other models are tested. The full dataset was divided into a training set and a testing set. Stratified sampling was used in the division of dataset. In this sampling process, 25 samples or 5/6 of each class were randomly selected and put into the training set. The remaining 5 samples or 1/6 were put into the testing set. Hence the proportions of each class are equal in both training and testing set. This division to training set and testing set was carried out three times randomly. The models were applied to these three random sets and the average performance was used to evaluate the performance of the models. Table 4 shows the experimental results.

Table 4 Classification of Gabor signatures

Method	Settings	Dataset 1		Dataset 2	
		Training	Testing	Training	Testing
NN		93.33%	94.67%	94.40%	90.67%
BP	10 neurons	96.27%	88.00%	96.53%	97.33%
K-Means	K=5	93.07%	94.67%	81.87%	76.00%
K-Means	K=10	96.80%	97.33%	94.93%	92.00%
KFLANN	$\delta=\sigma, \rho=0.8$	94.93%	96.00%	93.33%	92.00%
KFLANN	$\delta=\sigma, \rho=0.9$	100.00%	100.00%	98.93%	97.33%

The results again indicate that Gabor signatures are good signatures for learning and classifying objects. For non-noisy image set, KFLANN attained a superior testing performance over other methods. For noisy image set, BP and KFLANN with vigilance set at 0.9 showed the best testing performance. In conclusion, the KFLANN with a high vigilance value (0.9) produced the best performance among all models.

3.3.5 Discussions

An edge orientation detection algorithm based on Gabor filters was presented. A combination of smaller and larger receptive fields is used to retrieve an accurate edge region and accurate edge orientations. The edge orientation information is used for constructing a Gabor signature. Due to the properties of Gabor filtering, this method shows robustness in the presence of Gaussian noise. The signature proposed is economical and is useful in constrained recognition applications where the edge and edge orientation information play important roles. Such applications are numerous in the automated inspection industry where the background is simplified and the edge information can be a robust source for the detection applications.

3.4 Vergence

When vergence capability is not present in the vision system, binocular single vision is missing and a phenomenon called diplopia occurs. This causes the confusion of a visual target's direction. Binocular single vision, called fusion, as well as binocular depth perception, called stereopsis, is either eliminated or severely degraded [123]. The objective of vergence is to make the centers of the binocular visual input be the same point in space. Difficulties in vergence control include significant disparity variations in a complex background, requirements for real time performance and mechanisms to detect instable fixation positions. The retino-cortical mapping of visual information, the spatial organization of visual information in the visual cortex, and the disparity tuning properties of binocular neurons have inspired many computational vergence control models in computer vision research. Amongst the four sources of vergence, disparity estimation is perhaps the most widely applied. In disparity based vergence control systems, the correlation based, phase based, and cepstrum based disparity estimation methods can all be considered as equivalent to retrieving a certain type of correspondence between the image pair. As the saccade is modeled as monocular operation in the CogV system, the investigations into binocular vision are more on vergence control models. An analysis of the vergence geometry is provided here.

3.4.1 Vergence geometry

Considering the situation in Figure 38, two cameras initially fixated on a distant object. The object then moved nearer, rendering positional displacement in both images. In the right image, the position of the object moves to the left. In the left image the object moves to the right. This produces a displacement of $d/2$ at the center of each image. The vergence angle θ is the angle between the two lines of sights. To fixate on the object

again, both eyes need to be adjusted by an angle α so the vergence angle changes from θ to $2\alpha + \theta$. 2α corresponds to the disparity d in the image domain and can be derived from a disparity estimation method plus a mapping from image to angle.

Figure 38 Disparity and vergence

Assuming the object in Figure 38 moves from a depth D_1 to D_2 , and the baseline distance is B , the relationships between the vergence angles and depth can be expressed as:

$$D_1 \tan\left(\frac{\theta}{2}\right) = \frac{B}{2}$$

$$D_2 \tan\left(\alpha + \frac{\theta}{2}\right) = \frac{B}{2}$$

The vergence movement α can be derived in Equation (25).

$$\alpha = \arctan\left(\frac{B}{2D_2}\right) - \arctan\left(\frac{B}{2D_1}\right) \quad (25)$$

From Equation (25), if D_1 , the previous fixation point is fixed, the depth D_2 determines the desired vergence movement α . It can also be deduced that a smaller baseline B will produce a smaller vergence movement α . Thus a smaller baseline makes the vergence task easier because the disparity on the image is proportional to the desired vergence movement α and smaller disparity range reduces the computational load of disparity estimation algorithms.

The scope of our vergence system limits the depth ranges from about 0.5 to 5 meters. In our current platform if the fixation point shifts from a depth of $D_1=5\text{m}$ to $D_2=3\text{m}$ in the case of Figure 38, the derived vergence movement $\alpha=0.916^\circ$. This corresponds to a disparity of $d=2\alpha/(A/M)=39.87$ pixels. In the master-slave design of the CogV system, assuming the master camera fixates correctly on the near target, the disparity estimator on the slave image must cover at least 40 pixels of disparity to achieve a relevant estimated disparity. With a similar calculation, when $D_2=2\text{m}$, the disparity increases to 89.5 pixels. Vergence is more difficult when the system shifts from a certain fixation position to another with larger depth changes. This problem is especially crucial in active vision systems where the primary objective is to actively explore and interact with the environment.

3.4.2 Vertical disparity and tonic vergence

Besides horizontal disparity, vertical disparity is also produced when the fixation point changes. Neuroscience experiments showed that disparity-selective neurons from the primary visual cortex modulated firing rate over a wider range of horizontal disparity than vertical disparity [124]. This is natural because in the human vision system vertical disparity is usually much smaller than horizontal disparity. In artificial vision systems, vertical disparities arise due to perspective distortion and motor rotations. In the CogV system, vertical disparity is also estimated and corrected in a smaller range than the horizontal disparity.

For initial vergence where the cameras start from parallel orientation, a tonic angle can be added to the system to provide a default vergence range instead of the initial parallel setting. This tonic vergence setting can reduce the vergence movements of the system

when fixating on some objects for the first time. This angle can be very task dependant, relating to the application domains. For example, indoor mobile robots would have a different tonic setting than an outdoor aero robot. Other components of vergence can also be used to limit the disparity in a certain range. For example, accommodation can be used to infer the target distance and adjust the vergence angle to move nearer to the correct position. Motion cues can also be included for matching and tracking a dynamic target. These are separate models for enhancing vergence. Although potential improvements can be derived from these factors, it is beneficial to study the working ranges of disparity based vergence and develop models working in a large range through analysis of the hardware, the working environment and the system configurations.

3.4.3 Developed vergence control models

The following chapters present the vergence control models designed for the CogV system. These models are categorized into four types. Type-A model is a global disparity estimation model based on localized comparison of edge information in log polar space. This model works only in constrained scenes where objects in the view field are visible to both cameras and the background is simple. Type-B model is a pyramidal disparity energy model, designed using binocular filters bearing similar characteristics to the binocular neurons in the visual cortex. This model provides a possibility of utilizing the disparity energy model to near real time vergence control. Type-C model is a pyramidal area correlation method, which provides adequate performance. An extension of this model to color images using an adaptive color mask is also introduced. Type-D model is a pyramidal correlation model using the log polar images which describes the visual-cortical mapping. Both Type-A and Type-D models were based on the log polar visual-cortical mapping. The Type-D model was shown to provide a robust performance under

complex scenes with significant depth changes. This method signifies a major contribution by the research conducted in this thesis. Both quantitative and qualitative comparisons were done to evaluate the models studied. In the developed vergence control models, the processing is based on grayscale images unless otherwise specified.

4 Type-A Vergence: A Log Polar Vergence Control System

The visual-cortical mapping in the human vision system can be modeled by the log polar transformation [19]. This chapter is a focused examination of the extent at which vergence can be attained by the log-polar organization found in the visual cortex. A method using edge features in the log polar space was proposed for vergence control in a simplified environment with no occlusions. The vergence control is carried out by a disparity computational model in the log polar space, inspired by several discoveries related to the primary visual cortex. The binocular images were first transformed to the log polar domain. The Lateral Geniculate Nucleus (LGN) was simulated by center surround differences to produce edge features in the log polar space. The edge images were columnwise interlaced to simulate the ocular dominance columns in the human visual cortex. An XOR operation was performed between neighbouring columns of the interlaced image to determine the fusing and non-fusing edges. The non-fusing edges were used to estimate the global shift in the binocular image. With the cortical magnification by the log polar transformation, the global shift of non-fusing edges gave an estimation of the fovea disparity, which was used as an error signal for vergence control. This system has a robust performance under occlusion free situations and showed the possibility of constructing a simplified model for significant vergence control. In this chapter the rationale for the system implementation is illustrated and the

properties of this model are quantitatively studied to illustrate its performance. The difference between operating in the visual domain and the cortical domain are compared. The major limitation of this model is that it does not work in cluttered environment because the model assumes all the edges in the scene are visible to both cameras and this is not true in cluttered scenes. Occlusion may also degrade the system performance since occluded edges are only visible to one camera.

4.1 Log polar transformation

Through analyzing the anatomical and physiological experimental data, it was found that the visual-cortical mapping of the visual field to the surface of the striate cortex can be characterized as a logarithmic conformal mapping [14-19]. The log polar transformation has been used in both disparity estimation and vergence control [5, 59, 83]. In certain applications, the log polar transformation is utilized as a rotation invariant recognition application [125]. Assuming that the visual plane is a typical XY plane centering in the middle of the image with (x_0, y_0) as the origin, the log polar transformation can be represented by Equation (26) and (27).

$$\rho = \log_a r = \log_a \sqrt{(x - x_0)^2 + (y - y_0)^2} \quad (26)$$

$$\varphi = \arctan\left(\frac{y - y_0}{x - x_0}\right) \quad (27)$$

ρ is the logarithm of the distance between a pixel (x, y) to the origin or the image center (x_0, y_0) and φ is the polar angle of the point (x, y) . The logarithmic base a can be determined by the original image size and the transformed log polar image size [126]. Equation (26) can be rewritten in the form of Equation (28) where W defines a circle centering the fovea in the visual plane. The $\log_a$ component further implies that as the

point draws further from the fovea, the number of points that can define the discrete quantized ρ will increase.

$$\rho = \log_a W \quad (28)$$

$$\text{where } W = \sqrt{(x - x_0)^2 + (y - y_0)^2}$$

Figure 39 is an illustration of the relationship between points in the visual plane and those in the cortical plane. The blobs corresponding to a single cortical point in the visual plane becomes larger as the distance to the fovea increases.

Figure 39 An illustration of the relationship between the visual space and the log polar transformed cortical space

Between the two transformation spaces, there exists a circle around the centre of the visual plane where there is a one-to-one pixel correspondence with the cortical plane [126]. The region inward of the one-to-one circle has a property where each point in the visual plane corresponds to more than one point in the cortical plane. Therefore a large number of neurons process visual information from a small area of the visual field. For the region outward of the one-to-one circle, each point in the cortical plane corresponds to more than one point in the visual plane. One neuron in the peripheral region

processes visual information from a large area of the visual field, which can be seen through Figure 39. This is why we have a high visual acuity for the fovea vision. When two eyes fixate on a single object, the objects in the periphery of two eyes will not fuse perfectly, especially if they are out of Panum's fusional area (Section 2.1.2.4). There will also be varying disparities depending on the point of fixation, the position of the object in the periphery and the structure of the object. If the object lies in the periphery, the many-to-one correspondence of the visual-cortical space plane compensates the disparity by a factor proportional to the distance from the point of fixation. For the fovea region (inside the one-to-one circle), a one-to-many correspondence of the visual-cortical space results in a fovea magnification. For example, the circle at the fovea region in Figure 40 is enlarged to occupy about three quarters of the log polar image. A slightest shift in the visual plane at the fovea will lead to a magnified shift in the cortical plane. Thus the log polar transformation provides a way of concentrating on the fovea disparity by the fovea magnification and periphery diminishment. It is with these two properties that the proposed vergence control model leverages on to provide an efficient and effective means of binocular vergence.

Figure 40 Log polar transformation: (a) original image and (b) log polar image

The proposed vergence control model employs a backward log polar transformation, which is a log polar sampling on the input image. This mapping from cortical plane to

visual plane is pre-calculated to increase the processing speed. As shown in Figure 40, the disparity estimation is dominated by the fovea due to the cortical magnification. Images falling into the periphery receive smaller significance, but still contribute to the vergence estimation.

4.2 Central visual pathway and ocular dominance columns

Besides the log polar transform, another property of the human vision system significant to influence the design of the proposed vergence mechanism is the contralateral and ipsilateral information processing in the central visual pathway [1]. Each cortical hemisphere of the brain obtains information from both retinas gathered from retinal cells on the same hemispheric half. This happens as approximately half of the optic-fibers cross at the optic chiasm while the rest remain on the same hemisphere before entering the Lateral Geniculate Nucleus (LGN). These connections eventually reach the visual cortex [1]. As a result of this segregation, the left and right halves of the two eyes are transmitted separately.

Another interesting discovery is the synergistic arrangement of the fibres of each half as it reaches the visual cortex. The visual information from the left eye and right eye are interlaced together to form a striped pattern of alternating eye signals, known as ocular dominance columns [11]. The neurons in the primary visual cortex are organized to spatially preserve information within a local region so that localized processing can be performed within the localized receptive fields. The presence of these stripes of alternating dominance prompted an investigation of the possibility that this naturally occurring organization is instrumental for vergence control. Any disparity in the scene when viewed from the binocular system would usually result in a localized displacement

seen within neighboring ocular dominance columns. Thus, if a global view of the dominance columns could be optimized to yield a minimal disparity value, vergence would be indirectly attained.

4.3 The vergence control model

This section presents the vergence control model which produces a vergence error for the movements of cameras. The amplitude of this vergence error is not exactly equal to the true disparity at the fovea of the image center but through minimizing this vergence error, the disparity at the center of the image can be minimized.

4.3.1 The algorithm

The vergence control algorithm is shown in Figure 41. The two input images are initially mapped to the log polar space. The input images have resolution of 1000×1000 and the transformed log polar images have resolution of 200×180. Then LGN center-surround operator is applied on the log polar images for edge detection. Edge detection is applied to the log polar images, which unlike typical applications of image processing, in which this is usually performed in Cartesian or visual domain. A simple point contrast detector is used where the edge strength is produced by the difference between the center pixel and the average of its 8 neighbors. The resultant binocular pair of log polar edge images is interlaced columnwise to simulate the localized comparison between points from the left visual field and the right visual field. A localized XOR operation is performed on the interlaced image for each pixel (x,y) and its next horizontal neighbor $(x+1,y)$ where (x,y) is from the left image and $(x+1,y)$ is from the right image. The result of the XOR operation is an XOR map that indicates the non-matching edges from the two log polar edge images. The vergence error is produced from these non-matching pixels.

Figure 41 Cyclical flowchart for vergence control beginning with a view change

Assuming that the entire information incident from the visual space falls within the Panum's fusional area, the disparity estimation becomes analogous to a fusion process and all mismatched edges will contribute to the disparity signal. The following section formalizes the discussions made so far and presents the process of generating the horizontal vergence error. We proceed with a description of the notations.

Definition: Let L and R denote the left or right camera respectively and the number following the camera label denote the visual hemispheres 1 and 2 (left or right hemisphere respectively).

P_{L1} is the set of edge point pixels in the **left** visual hemisphere of the **left** camera which pass the XOR operator (which means the corresponding position in the **right** image is not an edge pixel)

P_{L1n} is the horizontal position of a P_{L1} pixel n in visual space with respect to the center of the image

C_{L1} is the number of pixels in P_{L1}

P_{R1} is the set of edge point pixels in the **left** visual hemisphere of the **right** camera which pass the XOR operator (which means the corresponding position in the **left** image is not an edge pixel)

P_{R1n} is the horizontal position of a P_{R1} pixel n in visual space with respect to the center of the image

C_{R1} is the number of pixels in P_{R1}

The information from the left half and the right half of the visual field is processed separately. The summed horizontal positions of pixels in P_{L1} and P_{R1} can be expressed as

$$pos_{L1} = \sum_{n=1}^{C_{L1}} P_{L1n}$$

$$pos_{R1} = \sum_{n=1}^{C_{R1}} P_{R1n}$$

The difference between the summed horizontal positions, normalized by the number of pixels, produces a global disparity for the hemisphere 1 (left hemisphere) of both cameras L and R .

$$disp_1 = \frac{pos_{L1} - pos_{R1}}{C_{L1} + C_{R1}}$$

Similar processing is done for the right visual hemispheres using $L2$ and $R2$, where all annotations are similarly defined.

$$pos_{L2} = \sum_{n=1}^{C_{L2}} P_{L2n}$$

$$pos_{R2} = \sum_{n=1}^{C_{R2}} P_{R2n}$$

$$disp_2 = \frac{pos_{L2} - pos_{R2}}{C_{L2} + C_{R2}}$$

Considering both hemispheres as important in determining the final displacement to minimize the global disparity, the two normalized disparities must be summed together to give the final disparity.

$$\begin{aligned}
Disparity_{pan} &= disp_1 + disp_2 \\
&= \frac{\sum_{n=1}^{C_{L1}} P_{L1n} - \sum_{n=1}^{C_{R1}} P_{R1n}}{C_{L1} + C_{R1}} + \frac{\sum_{n=1}^{C_{L2}} P_{L2n} - \sum_{n=1}^{C_{R2}} P_{R2n}}{C_{L2} + C_{R2}}
\end{aligned} \tag{29}$$

Equation (29) summarizing the disparity computations approximates the general behavior of displacement minimization. A threshold can be set such that the system assumes that vergence is achieved when disparity is below the threshold. Being a global method, the disparity produced is actually a global displacement of edge features in the image, giving the system a direction for correct vergence. The exact value of the disparity may not possess equal amplitude with the true disparity at the image center.

4.3.2 Dynamics of disparity estimation

An example is used to illustrate the vergence control process. In this example, the emphasis is on objects located near the fovea. Figure 42 shows a simulated sequence of frames taken from a pair of cameras, imaging a single black square.

Figure 42 Illustration of the algorithm

The left camera is the master eye and is deliberately set to fixate at the center of the square. The right camera is set as slave and the image captured is centered to the right of the square. Take the column Left and column Right 1 in Figure 42 as an example. This binocular pair of images goes through log polar transform and LGN edge detection. The log polar edge images are then interlaced and compared to provide the cortical map. A pixel-level XOR Boolean operation was used to determine the areas of mismatch. Each pixel in the left image is compared with the pixel at the same image coordinate in the right image. If they are similar, i.e., both are either edge points or non edge points, the corresponding pixel in the resultant cortical image is set to grey, which means these pixels don't contribute to the disparity. If the left pixel is an edge point and the right pixel is not, the position is set to black in the resultant cortical image. Similarly, if the right pixel is an edge point and the left pixel is not, this position is set to white in the resultant cortical image. The position difference of the black pixels and white pixels in the cortical map gives the vergence error. The camera can be moved by the vergence error to produce column Right 2 in Figure 42. This is repeated until the vergence error is within the pre-defined threshold. At this moment, the cortical map should either be all grey or the positions of black and white pixels are nearly equal to provide a low disparity.

According to Equation (29), when the position of the object in the right eye is to the left of the position of the object in the left eye, the $Disparity_{pan}$ will be positive because the disparity is defined as left minus right. Similarly, when the position of the object in the right eye is to the right of the position of the object in the left eye, the $Disparity_{pan}$ is negative. When the positions of the object in both eyes match, the $Disparity_{pan}$ will be zero. In general, there are four possible scenarios that the vergence system will cycle through. These can be classified into the following 4 cases.

Case 1: The square appears in the left visual field of the right eye as shown in column

Right 1 of Figure 42. In this situation:

$$C_{L1} = C_{L2} \text{ and } \sum_{n=1}^{C_{L1}} P_{L1n} = -\sum_{n=1}^{C_{L2}} P_{L2n} < 0$$

$$C_{R1} > 0, C_{R2} = 0 \text{ and } \sum_{n=1}^{C_{R1}} P_{R1n} < 0, \sum_{n=1}^{C_{R2}} P_{R2n} = 0$$

The disparity between the left and the right eye will be:

$$\begin{aligned} Disparity_{Pan} &= \frac{\sum_{n=1}^{C_{L1}} P_{L1n} - \sum_{n=1}^{C_{R1}} P_{R1n}}{C_{L1} + C_{R1}} + \frac{-\sum_{n=1}^{C_{L1}} P_{L1n} - 0}{C_{L1}} \\ &> \frac{\sum_{n=1}^{C_{L1}} P_{L1n} - \sum_{n=1}^{C_{R1}} P_{R1n}}{C_{L1} + C_{R1}} + \frac{-\sum_{n=1}^{C_{L1}} P_{L1n}}{C_{L1} + C_{R1}} = \frac{-\sum_{n=1}^{C_{R1}} P_{R1n}}{C_{L1} + C_{R1}} > 0 \end{aligned}$$

The slave (right) eye should pan to the left.

Case 2: Part of the square appears in the left visual field of the slave eye and the rest in the right visual field, shown in column Right 2 of Figure 42. Because the objects are near to the center of both eyes, it is assumed that there are approximately equal number of points on the four lines in the cortical map. So we have

$$C_{L1} = C_{R1} \text{ and } \sum_{n=1}^{C_{L1}} P_{L1n} - \sum_{n=1}^{C_{R1}} P_{R1n} > 0$$

$$C_{L2} = C_{R2} \text{ and } \sum_{n=1}^{C_{L2}} P_{L2n} - \sum_{n=1}^{C_{R2}} P_{R2n} > 0$$

$$Disparity_{Pan} = \frac{\sum_{n=1}^{C_{L1}} P_{L1n} - \sum_{n=1}^{C_{R1}} P_{R1n}}{C_{L1} + C_{R1}} + \frac{\sum_{n=1}^{C_{L2}} P_{L2n} - \sum_{n=1}^{C_{R2}} P_{R2n}}{C_{L2} + C_{R2}} > 0$$

In this case, the slave (right) camera should also pan to the left.

Case 3: When slave eye is also correctly foveating on the object, the two images will superimpose (assuming all information is from Panum's fusional area). As a consequence the left camera will not move.

$$Disparity_{Pan} = 0$$

Case 4: The square appears in the right visual field of the slave eye. This case is exactly the inverse of case 1 and 2 and the derivation of $Disparity_{Pan}$ is similar.

$$Disparity_{Pan} < 0$$

In this case, the slave (right) camera should move to the right in the image.

4.4 Experimental Results

4.4.1 Comparisons with Cartesian or visual space

The proposed method was compared with a variation of the architecture with LGN edge detection performed in the visual or Cartesian domain rather than the cortical or log polar domain. The flowchart in Figure 43 shows the algorithm for comparison.

Figure 43 The flowchart of the algorithm for comparison

The Cartesian domain method performs comparisons ignoring the log polar transform step in Figure 41. There is no mapping from visual space to cortical space. Every step is performed in the Cartesian domain.

In this experiment binocular image pairs captured from the CogV system were used. Several magnetic shapes were put on a white board. The board was 1.5 meters away from the system and image sizes of 490×362 pixels were obtained. Both the left and right cameras were initially fixated on the black circle located at the centre of the image

as shown in Figure 44. The right camera was subsequently panned manually at a step of 2.5° to obtain a series of images with different disparities from the left camera. 11 pairs of binocular images were generated.

Figure 44 Left and right images foveating on the same object

The proposed log polar domain method and the Cartesian domain method were applied to the 11 image pairs. The vergence error produced were plot against the horizontal disparity. Different edge detection threshold were tested to show the tolerance of both methods to the edge detection and noises in the images. The results are shown in Figure 45. The linearity of the disparity tuning curves in Figure 45 shows the possibility of utilizing the models for vergence control.

Figure 45 Vergence error estimation. Vertical axis shows the disparity estimated and horizontal axis shows the true disparity at a step of 2.5° .

It can be observed that for threshold 10 to 40, the Cartesian domain method overestimated the disparity. It was realized that although the camera verged on the center circle in Figure 44, the objects to the left and right side of the center has a positive disparity due to the perspective changes because the right camera had to pan to the left to verge on the circle. This can be seen from the interlaced edge image in Figure 46. The objects in the right image showed positive disparities compared to the objects in the left image. The Cartesian domain method is based on the average of the disparity in the whole image, thus it produced an average disparity (positive) for all object in the image. The log polar space method didn't overestimate the disparity for threshold 10 to

30. This shows the advantage of log polar transformation which magnifies the fovea of the image.

Figure 46 The interlaced edge image. Red color is from the left edge image. Green color is from the right edge image. Overlapped edges are shown in yellow.

When the edge threshold is larger (40), the proposed log polar domain method also tended to produce a positive disparity. The reason can be seen from Figure 47. When the threshold was increased to 40, the edges from the center object got broken while the edges from the surrounding objects were still consistent. Therefore the estimation of disparity preferred the surrounding objects which have a larger weight now. When the threshold increased to 80, Figure 47d showed that the edge information was lost and neither the Cartesian domain method nor the log polar domain method produced a reasonable estimation. In the vergence control system, the selection of threshold is critical and needs to be carefully dealt with. In the proposed model, the threshold was empirically set to 30.

Figure 47 The edge detection on log polar images with threshold (a) 10, (b) 20, (c) 40 and (d) 80

4.4.2 Repeatability of positioning

This experiment was conducted to examine the repeatability of the vergence control model. The system was configured as an autonomous saccade-vergence system that foveates on objects attached to the white board with the system set at 1.8m from the board. The configured system had the ability to saccade and cycle through each object at the unique position only once. The saccade positions were generated by the Watershed segmentation model presented in Section 3.2.1. The result is shown in Figure 48.

Figure 48 Vergence control results: the initial parallel image pair and image pairs after vergence control on an object. The dotted squares show the fovea of each image.

For each object on the wall, the system was permitted to perform vergence three times. The ground truth pixel in the right image corresponding to the center pixel of the left

image was manually selected. The difference between the obtained vergence position and the ground truth produces the vergence control error. The average vergence control error is shown in Table 5. The resolution of the binocular image was set at 1000×1000. One pixel in the image corresponds to about 0.05 degrees and the object sizes in the image are 50 to 100 pixels. The results in Table 5 indicate that the system is able to perform accurate vergence control with a worst case pan error of 0.8 degrees and tilt error of 0.13 degrees.

Table 5 Vergence control errors

Object	Average pan error (pixels)	Average tilt error (pixels)
1	1.33	2.33
2	4.67	2.33
3	3.33	2.67
4	3.33	1.67
5	2.33	1.67
6	16.00	0.67
7	9.33	1.67
8	9.00	1.00

4.4.3 Depth and distance estimation

In this experiment, two objects were used to evaluate the vergence performance. As shown in Figure 49a, two objects, a round container and an aluminum water-bottle were placed in front of the system. A deliberate effort was made to prevent ambiguity from occlusion. Figure 49b shows the resultant image sequences in initial and the vergence states, where the object at the center of the image is the intended target. The system successfully detected and performed saccades and vergence on the two objects.

Figure 49 (a) Configuration of the experiment: side view and top view (b) initial status and vergence on the objects

The distance between the objects and the system were calculated through triangulation. Table 6 shows that the system was able to estimate the depth and distance to the objects at accuracy above 90%.

Table 6 Depth and distance estimation

Object	True depth (m)	Estimated depth (m)	True cyclopean distance (m)	Estimated cyclopean distance (m)
Bottle	1.60	1.70	1.69	1.78
Box	1.00	0.95	1.06	1.01

4.5 Discussions

The vergence control model presented in this chapter is a biologically inspired technique. It utilizes concepts derived from the human vision system and models the visual-cortical mapping, LGN center surround processing and localized comparison by ocular dominance columns. The vergence error extracted is a global error with a higher sensitivity in fovea yet not completely ignoring the periphery. As such, it is not confused by similar objects in the scene, even when the objects are in close proximity to each

other or on the same horizontal axis. This can be seen from Figure 48 that the three circles didn't cause any confusion of the system.

The vergence control is fast due to the log polar transformation. All the processing operations have a complexity of $O(P \times Q)$ where $P \times Q$ is the log polar image size. The log polar map size is normally much smaller than the Cartesian image size so the vergence control is much faster than the methods that processing information in the visual plane.

However the vergence control model is not completely suitable for a cluttered environment as it uses a global edge based algorithm. The assumption that the information are from Panum's fusional area is not always true and the assumption that all visual information is visible to both eyes is not always true. Therefore the model only works in constrained environment where background is simple and objects are visible to both cameras. It was then necessary to progress onto a more encompassing model, giving rise to the model in Chapter 0.

5 Type-B Vergence: Vergence Control using Disparity Energy Neurons

The displacement of corresponding positions in a binocular image pair can be used to derive depth information [127]. Many biologically inspired computational models had previously been proposed to address the disparity estimation problem [45, 54, 73-74]. The disparity energy model has thus far been considered the most neurophysiologically faithful representation of the stereo mechanism in the early visual cortex [77-78, 128]. In this model, the processing units are the complex cells with binocular receptive fields which receive visual stimulus from the two input images. The positional difference or the phase difference between the binocular receptive fields determines the tuned disparity of the complex cell. Among a set of complex cells tuned to different disparities, the complex cell with strongest response determines the disparity in the visual stimulus. In this chapter, we focus on the phase shift disparity energy model which relies on the phase difference between the binocular receptive fields to estimate disparity. While the phase shift disparity energy model provides a faithful model for disparity estimation, the computational complexity of the model impedes the use of it in real time robotic systems because the tuned disparity is normally much smaller than the receptive field size. Due to the limited range of tuned disparities in the phase shift disparity energy model, the Coarse-to-Fine method was used to increase the range of tuned disparity and improve performance [54]. However, to accommodate a practical range of disparities,

the size of the receptive field increases extensively, making the processing impractical for vergence control applications. In this chapter a vergence control system using a pyramidal approach of the disparity energy model is presented. Through the hybridization of position shift and phase shift in a pyramidal image structure, it is possible to reduce the computational load of the original Coarse-to-Fine disparity energy model [54] by a significant order. The developed vergence control model utilizes compact weighted orientation pooling to provide reliable disparity. The disparity produced at the center of the binocular image is subsequently used for the vergence control in the CogV system. Through the pyramidal image structure, the disparity energy model was finally plausible for actual near real time vergence control in the CogV platform. Vertical disparities were also compensated to provide more accurate vergence control for peripheral fixation positions. Presented here are also the theories and results of the vergence control, including a systematic presentation of both the qualitative and quantitative results. The proposed model was also compared to a relevant disparity energy model based vergence control method [47] in the literature and showed more robust performance.

5.1 The disparity energy model

The basic components of the disparity energy model are the binocular simple cells and complex cells. The binocular simple cells have receptive fields positioned on the two retinas which can be described by Gabor functions. The response of a binocular simple cell is the summation of the responses from the binocular receptive fields. In the disparity energy model, the complex cells receive inputs from a quadrature pair of simple cells and the response of a complex cell is modeled as the sum of squared responses of this quadrature pair of simple cells. In this section, the construction of

simple cells and complex cells and how these are related to disparity estimation are explained.

5.1.1 Gabor filters and disparity energy cells

The receptive fields of simple cells in the primary visual cortex can be modeled as Gabor functions [13, 80], shown in Equation (30). Gabor filter [129] is a robust feature retrieval tool and have been used in many computer vision applications such as face recognition [130], feature retrieval [131], stereo vision [132] and vision based navigation [133].

$$G(x, y, \varphi) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \cos(\omega x + \varphi) \quad (30)$$

$$x' = (x - x_0) \cos(\theta) + (y - y_0) \sin(\theta)$$

$$y' = -(x - x_0) \sin(\theta) + (y - y_0) \cos(\theta)$$

Gabor filters can be considered as 2 dimensional Gaussian filters modulated by a cosine function in one dimension. The parameter θ determines its orientation, which is the clockwise rotation of the filter. In Equation (30), $\frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right)$ is a Gaussian envelope. σ_x and σ_y are the Gaussian widths in the x and y axis respectively. The aspect ratio γ of the receptive field is given by $\gamma = \sigma_x/\sigma_y$. The function $\omega = 2\pi f = 2\pi/\lambda$ determines the spatial frequency f and period λ of the filter in the x axis. φ is the spatial phase of the filter along the x axis. The parameter σ_x and λ are related, where the ratio σ_x/λ determines the spatial-frequency bandwidth of the filter. This ratio varies in a limited range of 0.4 to 0.9 which incidentally corresponds to 2 to 5 excitatory and inhibitory stripe zones in a receptive field [80, 134]. In the proposed vergence control model, $\sigma_x/\lambda = 0.5$. When the excitatory and inhibitory stripe zones become wider (λ increases), the Gaussian envelope size increases proportionately (σ_x increases).

Since binocular simple cells receive input from both the left and the right visual inputs, we begin by letting the left receptive field be modeled as $f_l = G(x, y, \varphi_l)$ and the right receptive field be $f_r = G(x, y, \varphi_r)$. Let I_l and I_r be the input image from the left and right eye respectively. The response r_s of a binocular simple cell can be expressed as Equation (31) where r_s is the summation of the individual left and right responses. The simple cell responses are produced by convolution between the receptive profile as a spatial filter and the visual input as an image.

$$r_s = r_l + r_r = f_l * I_l + f_r * I_r \quad (31)$$

where $*$ is the convolution operator

From physiological evidence, binocular complex cells can be modeled as the summation of squared responses of a quadrature pair of simple cells [78]. The quadrature counterpart of f_l and f_r are two filters with a phase difference of $\frac{\pi}{2}$: $f_{lq} = G\left(x, y, \varphi_l + \frac{\pi}{2}\right)$ and $f_{rq} = G\left(x, y, \varphi_r + \frac{\pi}{2}\right)$. With the $\frac{\pi}{2}$ difference, a quadrature pair actually forms the real and imaginary component in the complex plane if the Gabor filters are modeled as complex Gabor filters as those in Section 2.2.2.2. The response of this quadrature counterpart r_q is derived using Equation (32).

$$r_{sq} = r_{lq} + r_{rq} = f_{lq} * I_l + f_{rq} * I_r \quad (32)$$

The complex cell response is therefore the sum of squared responses of the quadrature pair of simple cells, shown in Equation (33).

$$r = r_s^2 + r_{sq}^2 = (f_l * I_l + f_r * I_r)^2 + (f_{lq} * I_l + f_{rq} * I_r)^2 \quad (33)$$

Each complex cell is tuned to a specific disparity determined by the position-shift or the phase-shift between the left and right receptive fields. To obtain the disparity, complex cells tuned to a range of disparities are combined to form a disparity energy model.

There are primarily two types of disparity energy models, namely the position-shift model [73, 135] and the phase-shift model [78].

5.1.2 The position-shift model and the phase-shift model

Equation (34) and (35) define the receptive fields of the cells in the disparity energy model.

$$f_l = G(x, y, \varphi_l) \text{ and } f_r = G(x + d, y, \varphi_r) \quad (34)$$

$$f_{lq} = G\left(x, y, \varphi_l + \frac{\pi}{2}\right) \text{ and } f_{rq} = G\left(x + d, y, \varphi_r + \frac{\pi}{2}\right) \quad (35)$$

The position-shift model assumes that the left and right receptive fields of a complex cell have the same receptive profile ($\varphi_l = \varphi_r$) but can be centered at different positions ($d \neq 0$) in the left and right retina. The shift between the spatial locations of the left and right receptive field centers is the preferred disparity of the complex cell. By defining a group of n complex cells with different position shifts, the disparity d_c existing in the visual inputs can be determined by the complex cell c with the strongest response.

$$d_c = \operatorname{argmax}_d \{r(d)\}, \quad d \in [d_{min}, d_{max}] \quad (36)$$

Unlike the position-shift model, the receptive field centers in the phase-shift model are the same ($d = 0$) in the left and right retina. However, the phases of the left and right profiles are different ($\Delta\varphi = \varphi_l - \varphi_r \neq 0$) and this determines the preferred disparity of the complex cell. By defining a group of n complex cells with different phase shifts, the disparity d_c existing in the visual inputs can be determined by the complex cell c with the strongest response. The vergence control model proposed in this paper uses the phase shift model to represent the basic processing unit as it has been shown that for relative small disparities, the population response generated by the phase shift model is more reliable than that generated by the position shift model [54].

Assuming $\theta = 0$ for the detection of horizontal disparity D , after taking the Taylor series expansion of the simple cell receptive functions, the complex cell response of the phase-shift model can be re-written in the form of Equation (37). A complete derivation from this expansion is provided in Appendix A. Note that the disparity D is defined as the horizontal image coordinate of a point in the left image minus the horizontal image coordinate of the corresponding point in the right image.

$$r = 4A^2 \cos^2\left(\frac{\Delta\varphi - \omega D}{2}\right) + 4A \frac{D}{\sigma_x} B \cos\left(\frac{\Delta\varphi - \omega D}{2}\right) \cos\left(\alpha - \beta - \frac{\Delta\varphi - \omega D}{2}\right) + \frac{D^2}{\sigma_x^2} B^2 \quad (37)$$

In Equation (37), the first term is normally larger than the second and third terms (see Appendix A). Keeping only the first term, an approximation of the complex cell response is expressed in Equation (38).

$$r = 4A^2 \cos^2\left(\frac{\Delta\varphi - \omega D}{2}\right) \quad (38)$$

From Equation (38), the maximum response occurs when the cosine factor is 1 implying that $\Delta\varphi - \omega D = 0$. Hence the preferred disparity D of a complex cell is determined by the intrinsic property of the cell, which is the phase difference between the left and right receptive field profiles. When the stimulus disparity is consistent with the preferred disparity of a complex cell, the cell will response strongly.

$$D_{pref}^{pha} \approx \frac{\Delta\varphi}{\omega} = \frac{\varphi_r - \varphi_l}{\omega} \quad (39)$$

Due to the periodicity of phase, the phase-shift disparity energy model produces disparities in a limited range, which typically corresponds to half the spatial period of the cosine term. By defining a group of n complex cells with different phase differences $\Delta\varphi_m$ ($m \in [1, n]$), the foveal disparity d_c existing in the visual inputs can be determined by the complex cell c producing the strongest response.

$$d_c = \frac{\operatorname{argmax}\{r(\Delta\varphi)\}}{\omega}, \quad \Delta\varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \quad (40)$$

The process of filtering an image with a complex cell tuned to disparity +2 is shown in Figure 50. This cell produces a response map from the two input images. If a group of complex cells are defined tuned to a certain range of disparity, the response maps produced from this group of cells can be used to determine the disparity for each position in the image.

Figure 50 A complex cell example using the phase shift disparity energy model: $\varphi_l=0$, $\varphi_r=\pi/2$, $\omega=\pi/4$. The preferred disparity is +2. The input image size is 200×200 and the receptive field size was 29×29. The center of the binocular image pair is expected to converge to a disparity of 0.

5.1.3 Pooling and population responses

Complex cells are arranged in unique groupings to produce maximal response when stimulus with the corresponding orientation is encountered. While the objective of our disparity estimation system primarily leverages on the vertical responses, single

orientation detection methods are prone and susceptible to noise. It is possible to improve detection performance through pooling of information from neighboring orientations. Similarly, in the spatial and scale dimensions, a synonymous expansion of pooling can be introduced. A population of pooled responses is therefore expected to increase the reliability for disparity estimation. Three types of pooling can be applied in the disparity energy model, namely spatial pooling, orientation pooling and scale pooling [54].

Spatial pooling of complex cell responses within a small neighborhood has been shown to significantly improve the quality of computed disparity maps [54]. This can be considered as a smoothing process to remove spike noises in the disparity maps and distribute information to neighboring cells.

Orientation pooling was also shown to be effective in estimating disparity. Due to the oriented nature of the Gabor filters, for the oriented stimulus, the filter that has a matching orientation is most accurate in estimating the disparity. For non-oriented stimulus such as complex textures, pooling through a range of orientations can capture more information and make the output more reliable. In the proposed model, orientation pooling within a narrow range was used. The filter orientations were constrained to be within $\pm 30^\circ$ with respect to the vertical orientation. Since orientations deviating more than 45° from the expected orientation tend to be affected by the change in the orthogonal orientation, the larger orientation angles were not considered.

Figure 51 shows the orientation pooling and spatial pooling processes. The input images were filtered by the disparity units with different tuned disparities (indicated by $\Delta\varphi$).

Each disparity unit contains a group of complex cells with the same tuning property ($\Delta\varphi$) but different orientations ($\theta_1 - \theta_n$). Orientation pooling combines the oriented response maps together by summation of responses from different orientations. Spatial pooling applies a Gaussian filter to the pooled response maps. Final disparity estimation is through detecting the maximum responses from the resulting response maps.

Figure 51 The orientation pooling and spatial pooling process

Another type of pooling, scale pooling, was performed to widen the tuned disparity range to address the narrow range problem of phase-shift disparity energy cells due to the periodicity of phases. As the disparity energy model can only detect disparities within about half of its period, different scales of complex cells were used for the Coarse-to-Fine processing to enlarge the tuned disparity range [54]. Larger scale cells were used for detecting larger disparity which is propagated to smaller cells for fine tuning of disparity. However, the scale pooling requires a significantly large computational resource because of the large receptive field size to detect large disparity. In this chapter an alternative method is adopted whereby a pyramidal image structure is

utilized instead of different scales of cells. In this way, the computation at different layers in the pyramid is kept constant, yielding a faster computation than the filter-scale-change approach. The equivalence of the two methods in estimating disparity will be shown through experimental evidence. It will be shown that this new approach is suitable for implementation into near real time vergence control.

5.2 The pyramidal model for disparity and vergence

Coarse-to-Fine methods have been applied in areas of image processing and computer vision [53]. In the Coarse-to-Fine disparity energy model [54], the disparity estimation begins with large scale disparity energy cells tuned to a large range of disparity. This provides an estimation of the disparity at a coarser scale and larger step. The disparity is propagated to the next finer level and produces the position-shift of the receptive field. The model reduces the size of the receptive field and estimates the disparity at the position shifted by the coarser disparity. This continues until the finest scale is reached. The estimated disparity is accumulated through position shift of receptive fields at different levels. This Coarse-to-Fine method is a hybridization of the position-shift and phase-shift disparity energy model using different scales of disparity energy cells. The Coarse-to-Fine method is effective but computationally expensive during the employment of large receptive field sizes.

Figure 52 The image pyramid and fixed tuning spatial filters

This section introduces a pyramid image structure that reduces the computational complexity for disparity calculations. As illustrated in Figure 52, a pyramid structure with a fixed cell size is used to produce the similar effect of the Coarse-to-Fine approach. Instead of altering the cell sizes, the pyramidal model modifies the input image I_l and I_r at each level of the pyramid. The complex cell response at each level is calculated using Equation (41).

$$r = (f_l * I_l^n + f_r * I_r^n)^2 + (f_{lq} * I_l^n + f_{rq} * I_r^n)^2 \quad (41)$$

where I_l^n and I_r^n are the scaled image with size $\frac{M}{R^{L-n}} \times \frac{N}{R^{L-n}}$ at level n ($n \in [1, L]$), R is the scaling ratio between successive levels and $M \times N$ is the original image size

Computation at different layers of the model is less intensive than the Coarse-to-Fine model [54]. The comparison between these two methods is illustrated in Figure 53. In the next sections, both models are eventually shown to produce an equivalent and effective estimation of disparities in the image.

Figure 53 (a) The pyramidal method where the receptive field size is fixed and the images are sub-sampled to form a pyramid and (b) the Coarse-to-Fine method [54] where the image does not change and the filters are scaled.

5.2.1 Orientation and spatial pooling

5.2.1.1 Compact orientation pooling

As the human vision system has a natural horizontal alignment of the two eyes, it is expected that the horizontal tuning range of disparities is larger than the vertical. Experiments on mammals have shown that for many disparity tuned neurons in the primary visual cortex, the range of encoding is larger for horizontal than vertical disparities [124]. Due to the oriented nature of the complex cells, a large portion of complex cells are tuned to both horizontal and vertical disparity [136]. Therefore the presence of vertical disparity may affect the tuning to horizontal disparity if orientation pooling is applied. The tuning of horizontal disparity in the presence of vertical disparity was shown through a simulation using random dot stereogram as visual input. A group of complex cells were constructed, tuned to a disparity range of $[-4, 4]$. The orientation of the complex cells was set to $\pi/4$ from vertical. Random dot stereogram pairs with

horizontal disparity of 0 and vertical disparity of $[-4, 4]$ were used as input to the complex cells and the response curves were plotted against the cell's tuned disparity for each stereogram pair, shown in Figure 54.

Figure 54 Responses of a group of complex cells with orientation $\vartheta=\pi/4$.

The results indicate that for vertical disparity = 0, the horizontal disparity is correctly detected as a peak response at 0. While for other vertical disparities, the detection of horizontal disparity was biased by the vertical disparity. As the vertical disparity grew larger, the bias also became larger. This $\pi/4$ orientation cell is equally tuned to both horizontal and vertical disparity and thus a non-zero vertical disparity will affect the horizontal disparity tuning. A test on the responses, when horizontal disparity is not 0, was also conducted. The results similarly showed that vertical disparity does affect the estimation of horizontal disparity for a cell having an oblique orientation.

Therefore when disparities along a certain orientation are to be detected, the orientation pooling of the disparity energy model should be more compact and closer to that orientation. In vergence control, the horizontal component is most significant as the binocular pair is horizontally aligned. As such to detect the horizontal disparity, orientation pooling can focus on the orientations in proximity to the vertical. The five orientations selected for the vergence control model were $\theta \in \left[-\frac{\pi}{6}, -\frac{\pi}{12}, 0, +\frac{\pi}{12}, +\frac{\pi}{6}\right]$, shown in Figure 55.

Figure 55 The orientation pooling

5.2.1.2 Weighted orientation pooling

Instead of a pure pooling by summation of the orientation responses, which were used in the Coarse-to-Fine model [54], an orientation pooling using a weighted combination of the complex cell responses was used. Each complex cell response is weighted by a factor $\cos \theta$ to indicate its contribution to the horizontal disparity estimation, allowing cells that have an orientation nearer to the perpendicular orientation of the disparity to contribute more to the estimation.

$$r = \sum_{\theta} r(\theta) \cos(\theta), \quad \theta \in \left[-\frac{\pi}{6}, -\frac{\pi}{12}, 0, +\frac{\pi}{12}, +\frac{\pi}{6}\right] \quad (42)$$

5.2.1.3 Spatial pooling

The spatial pooling is realized through filtering the orientation pooled response map using a Gaussian function, similar to the original Coarse-to-Fine model [54]. The output is the weighted sum of the responses in a local neighbourhood.

$$r = r * G \quad (43)$$

where G is a 2D Gaussian filter

In the proposed model, 5 pools of complex cells were used. Each pool of complex cells is tuned to one disparity with the disparity range covers half a period of the cosine term of the receptive field function, centered on disparity 0. Each pool contains 5 complex cells corresponding to 5 orientations. The eventual composite response of a pool of complex cells is the Gaussian smoothed summation of the summation of the 5 weighted complex cell responses.

5.2.2 Multilevel disparity estimation

Figure 56 The flowchart of the pyramidal disparity energy model

Figure 56 shows the process of the disparity energy filtering using the pyramid image structure. The parameters of the disparity energy model used are shown in Table 7 . In order to compare the performance of this algorithm, a Coarse-to-Fine disparity energy model [54] with equivalent parameterizations was also implemented. The pyramidal model contains 3 levels and the Coarse-to-Fine model has complex cells of 3 scales. The pyramidal model possessed a scaling ratio of 2.0 between images at successive levels

while the Coarse-to-Fine model maintained a similar scaling ratio of 2.0 in the spatial frequency of the complex cells of successive levels.

Table 7 Parameters of the proposed pyramidal model and the Coarse-to-Fine disparity energy model [54]

Parameters		Pyramidal model	Coarse-to-Fine model [54]
Simple cell parameters	$\omega=2\pi f=2\pi/\lambda$	$\pi/4$	$\pi/16, \pi/8, \pi/4$
	Size	29	$29 \times 4, 29 \times 2, 29$
	σ_x	4	$4 \times 4, 4 \times 2, 4$
	σ_y	8	$8 \times 4, 8 \times 2, 8$
	θ	$-\pi/6, -\pi/12, 0, \pi/12, \pi/6$	
Spatial pooling parameters	Pooling size	11×11	
	σ	4	
Number of pools/disparities of complex cells			5

The pyramidal model maintained 5 pools of complex cells with each pool containing a group of complex cells tuned to the same disparity. The phases of φ_l for the 5 pools were evenly distributed across the range $\left[-\frac{6\pi}{8}, +\frac{2\pi}{8}\right]$. The phase difference of the left and right receptive field profile for the 5 pools were evenly distributed across the range $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$ corresponding to the disparities of $[-2, -1, 0, +1, +2]$. The implemented Coarse-to-Fine model also included 5 pools of complex cells corresponding to the 5 preferred disparities for each scale. These 5 disparities were evenly distributed across the range $[-8, +8]$, $[-4, +4]$ and $[-2, +2]$ respectively. Both the pyramidal and Coarse-to-Fine models maintain a group of complex cells that were functionally similar. The range of preferred disparities in the pyramidal model at each level was designed to be analogous to the system designed under the Coarse-to-Fine model.

5.2.3 Functional equivalence

The functional equivalence of the pyramidal model and the Coarse-to-Fine model were compared using disparity estimation results. The number of pools was set to 9 in this experiment to facilitate dense disparity estimation and compare the accuracy. A 200×200 random dot stereogram experiment was conducted with both methodologies and the results obtained are shown in Figure 57.

Figure 57 Disparity estimation results on a 200×200 random dot stereogram with dot density 0.5. The last column shows a mesh plot of the disparity map generated. Level 1 and level 2 results of the pyramidal model was $\frac{1}{4}$ and $\frac{1}{2}$ sizes of level 3 and were scaled to the level 3 size for comparison purpose.

At level 3 (the finest level), both methods were able to construct the disparity map. The resultant disparity map at level 3 of the pyramidal model and filter scale 3 of the Coarse-to-Fine method are visually compatible. At level 1, both models produce a poorer initial estimate of disparity, but both methods were capable of recovering the disparity through the layered processing. In terms of processing costs, the computational processing in level 1 of the pyramidal model costs less than the filter scale 1 of the

Coarse-to-Fine model. This is because the pyramidal model required both the filter size and input image size to be a quarter of the Coarse-to-Fine model's required resolution. Qualitatively, the initial computational estimations at the coarser levels may propagate through to the finer levels, resulting in slight defects in the final results. This can be seen from the stripe patterns at level 3 or filter scale 3. This can be improved through the use of a smaller scaling ratio such as $\sqrt{2}$ used in [54] and a larger number of scales. A scaling ratio of 2 was used in our experiments to improve the speed of processing and widen the tuned disparity range with the same number of levels. Computationally, it is more viable to operate image scaling rather than filter scaling as the former reduces the size of the data to be processed. With this reduced computational complexity, a vergence control system was implemented.

5.2.4 Vergence control

There have been many successful implementations of stereoscopic systems which were able to utilize optical geometry and disparity to determine object distances and these have undoubtedly contributed significantly to the research area of scene understanding [137]. Since vergence control for a binocular pair requires only the disparity at the image center, the pyramidal method was applied to the image center. Orientation pooling and spatial pooling were maintained and the filtering was restricted to the Gaussian neighbourhood of the spatial pooling filter centered at the image center. The center disparity estimated was used for controlling the cameras so that the image centers fuse, inherently attaining a successful vergence.

To ascertain control, one camera is dedicatedly assigned as the master camera and the other as the slave camera. The master camera is controlled to fixate on an object or an

area in the image and the slave camera that is initiated on a parallel axis with the master camera moves by the same amount to the same direction with the master camera. The vergence control is subsequently activated and the slave camera converges towards the master camera according to the image disparities computed at different levels of the pyramid. There is evidence that such master-slave configuration exists in the biological vision system [138] and the principle has also been used by existing vergence control systems.

Another implication of the experiment in Figure 54 is that vertical disparity must be corrected to improve the reliability of horizontal disparity estimation. In vergence control applications, vertical disparity arises as a result of image distortion and imperfect motor rotations. In the developed vergence control model, vertical disparity is also corrected. Vertical disparity is detected by rotating the whole disparity energy model by 90°. As vertical disparity is typically smaller than horizontal disparity, the disparity energy model was simplified to detect a disparity range that covers only ¼ of the period of the cosine term. The detectable vertical disparity range is therefore only half of the horizontal disparity range. The detected vertical disparity was corrected from coarser to finer level simultaneously with the horizontal vergence control.

The vergence control process undergoes the same vergence flow over the various levels of image granularity. The control goes from the coarsest to the finest level. At each level a proportional control rule, shown in Equation (44), is used for vergence control.

$$M = -K P d_t \quad (44)$$

K is the gain of the controller. It can be set to a lower value for slower vergence and higher for faster vergence but may introduce overshooting. K is set to 0.8 in this

vergence control model. P is a constant to convert the image domain disparity to the pan-tilt control signal. d_t is the estimated disparity at time t . The flowchart of the vergence control process is shown in Figure 58.

Figure 58 The flowchart of vergence control

At each granular level, vergence is pursued until both the horizontal and vertical disparity at the image center is less than a threshold (± 1 pixel was used here). Upon fulfilling each threshold requirement, the system progresses onto the next finer level and the same controlling process ensues. This continues until the finest level is reached and a high resolution vergence is attained. At each level, the vergence control model detects oscillation status by counting the number of opposite disparities in consecutive disparity estimation. Here 3 accumulated times of opposite disparities were used to define an oscillation situation. If oscillation occurs at the top coarser levels, the vergence control moves to the next finer level. However, if oscillation occurs at the finest level,

the status of vergence is indicated as unstable and the vergence control terminates. The final state of the vergence control is either verged or the system oscillates.

5.3 Experimental results

In the previous section, qualitative results showed that the proposed pyramidal model and the Coarse-to-Fine model [54] have compatible performance in disparity estimation. In this section, two sets of experimental results are presented. The first set presents the results of vergence control in a lab environment to illustrate the effectiveness of the developed model. Vertical disparity correction is shown to be necessary in the context of vergence control. The second set is a quantitative comparison between the pyramidal model and the Coarse-to-Fine model to determine if the two methods differed in accuracy and computational complexity. As such, a tabular comparison of the two models on the random dot experiment is included to appreciate the quantitative comparisons with ground truth. A quantitative understanding of the magnitude of time savings between the pyramidal method and the Coarse-to-Fine method is also presented to support the use of the pyramidal method for real time applications.

5.3.1 Vergence control and depth estimation

To successfully verge on a point, the disparity at this point should be within the disparity range of the disparity energy model. In our system, a 3 level pyramidal model was used to capture a maximum disparity of $[-14,14]$. This is calculated from the accumulative maximum disparity at 3 consecutive levels from Equation (45).

$$D_{max} = D \sum_{n=0}^{L-1} s^n \quad (45)$$

D is the maximum tuned disparity at each level,

s is the scaling ratio between successive levels,

and L is the number of levels

The two cameras capture 1000×1000 images corresponding to angle of view of 46°×46°. The images were sub-sampled to 200×200 resolution for the vergence processing. So the maximum disparity that can be handled by the system is about 3.22°. This corresponds to the horizontal disparity of about 24cm at a distance of 4m. The baseline between the two cameras is 24cm. This means the system can verge on objects 4m away from the platform, starting from a parallel configuration. An initial default vergence angle of 2.5° is set for the binocular vision system to simulate tonic vergence [9] so that theoretically the system can verge on objects at a distance of 2m away. Considering the human vision system which has a baseline of about 6cm, the working range of the developed system (if scaled to 6cm baseline) actually corresponds to objects at distance of 0.5m away, which is the within reaching distance of the human vision system. This is the distance where the human stereoscopic vision is most robust and is important for visual perception and interaction with the environment.

Figure 59 shows the results of vergence control in a lab environment. The fixation positions selected were 2 to 7 meters away from the system. The results showed that the system successfully verged on objects in this cluttered environment, with no oscillations detected.

Figure 59 The vergence control results: binocular images are shown in pairs and the gridded box indicates the center of image

To verify the effectiveness of vergence control, depth information of the fixated objects in Figure 59 were derived using the motor configurations in vergence state. The system performs vergence on each object three times and the average depth estimation results are shown in Table 8. The system showed a good performance for depth estimation with accuracy above 85%. The images for objects Box 2, Plant, Chair invoke a larger estimation error because they contain backgrounds with significant depth changes. The microphone stand produced a large estimation error as it was to the far left of the binocular system. The system was required to pan more than 60° leftwards and this causes image distortions and imperfect rotations on the fixed baseline system.

Table 8 Depth estimation results

Object	True depth (m)	Average estimated depth (m)	Average error	Standard deviation of error
Box 1	2.80	3.02	10.30%	5.76%
Poster	4.50	4.57	1.50%	2.02%
Box 2	2.80	3.21	14.57%	1.75%
Plant	7.00	5.95	15.02%	2.13%
Chair	3.30	3.56	15.05%	10.92%
Window	3.30	3.30	4.73%	1.96%
Projector	1.80	1.73	3.79%	2.92%
Door	3.80	3.56	6.42%	2.16%
Monitor	4.00	4.27	6.67%	2.62%
Box 3	2.30	2.30	2.71%	1.34%
Box 4	1.35	1.44	6.69%	5.84%
Mic stand	2.00	2.22	13.22%	8.85%

As the disparity energy model relies heavily on the phase information in a local receptive field, there exist situations where the model will not function well. Figure 60 illustrates two such cases. The first case illustrates the confusion caused by objects with similar appearances. The perspective changes made the right mug in the right image have a similar background with the left mug in the left image. As a result the right camera was directed to the right mug. The second case shows the over-biased contribution from the background where the targeted foreground is too small in the receptive field. The background now contributed a large proportion to the disparity estimation and directed the right camera to the background instead of foreground. This is the common problem of window based stereo matching methods where the assumption is the visual information within the window (receptive field of disparity energy neurons) is from the same depth but this is not always true in cluttered environment.

Figure 60 Failure cases of vergence control

5.3.2 Vertical disparity minimization

When the system tilts the cameras and pans to the periphery, vertical disparity always arises due to the distortion of images and the hardware limitations. It is necessary to correct the vertical disparity because it may affect the accuracy of horizontal vergence control. Figure 61 shows a case when we apply the vergence control with and without vertical disparity correction. Figure 61b shows that the vertical disparity actually biases the horizontal disparity.

Figure 61 (a) Vergence control with vertical disparity correction and (b) vergence control without vertical disparity correction

Figure 62 shows the process of verging on an object with both horizontal and vertical vergence control. The developed vergence control model shows robust performance for correcting both horizontal and vertical disparity.

Figure 62 The vergence control process. The frame numbers just indicate the successive sequence of images, not the true video frame number

5.3.3 Accuracy and computational complexity of disparity estimation

In order to quantitatively compare the results attained through vergence, a ground truth comparison was made with the calculated disparities. The random dot stereogram in Figure 57 was used to assess the accuracy of the pyramidal model and the Coarse-to-Fine model. A tolerance range was set such that if the absolute difference between the estimated disparity and the ground truth disparity was less than or equal to the tolerance, the disparity estimated would be considered correct. The boundary area of pixel width 15 (half the simple cell receptive field width) in the image was not considered to avoid bias due to lacking of information at boundaries. The results in Table 9 indicate that the pyramidal model achieves a slightly better result than the Coarse-to-Fine model. This result cannot be conclusive as to the statistical significance of this effectiveness between the methods. However, through several similar experiments of the quantifiable nature, the pyramidal model is shown to be at least compatible to the Coarse-to-Fine method.

Table 9 The accuracy of disparity estimation for the 200×200 random dot stereogram

Tolerance (in pixels)	Accuracy of disparity estimation	
	Pyramidal model	Coarse-to-Fine model [54]
0.5	89.40%	84.53%
1	95.28%	93.50%

Once the accuracy of the pyramidal model has been ascertained, the question then arises about the computational complexity for the pyramidal model and the potential reduction in processing time. Assuming the input image size is $M \times N$ and the receptive field size of the complex cells is $R \times R$. The complexity of filtering the input image using one complex cell is $O(M \times N \times R^2)$. For the pyramidal model, given P pools of complex cells where each pool corresponds to 1 disparity value containing A cells of A orientations with the size of spatial pooling of $S \times S$. Then the complexity of the disparity energy model for the finest level is $O(M \times N \times R^2 \times P \times A \times S^2)$. If there are L levels in the pyramid, then the total complexity is

$$\begin{aligned}
 & O\left(\left(\frac{1}{4^0} + \frac{1}{4^1} + \dots + \frac{1}{4^{L-1}}\right) \times M \times N \times R^2 \times P \times A \times S^2\right) \\
 & = O\left(\left(\frac{4}{3} - \frac{1}{3 \times 4^{L-1}}\right) \times M \times N \times R^2 \times P \times A \times S^2\right)
 \end{aligned} \tag{46}$$

The computational complexity is only $O(M \times N \times R^2 \times P \times A \times S^2)$ in the worst case. However, for the Coarse-to-Fine method, assuming a constant image size throughout and processed with a corresponding scaling of the filter size, then the total complexity for L scales is

$$\begin{aligned}
 & O\left(\frac{4^L - 1}{3} \times M \times N \times R^2 \times P \times A \times S^2\right) \\
 & = O(4^L \times M \times N \times R^2 \times P \times A \times S^2)
 \end{aligned} \tag{47}$$

This translates to an exponential increase of computational time for each additional scale. In comparison, the pyramidal approach reflects a linear complexity. In the

experimental setting, the Coarse-to-Fine model is expected to require $\frac{4^2+2^2+1}{0.25^2+0.5^2+1} = 16$ times the processing time of the pyramidal model. The computational time of the pyramidal model and the Coarse-to-Fine model are compared using the random dot stereogram and a pair of lab images captured by the CogV system. The two models were carefully implemented in MATLAB to bear architectural equivalence. The results were obtained under the processing platform of a 2.66GHz Intel CPU and presented in Table 10. Consistent to the expected improvements in speed, the pyramidal model reduces the computational time by a factor of about 10. In actual vergence control where the foveal region is the only position of processing, the computational time is further reduced. In the given scenario of the lab experiments, solely processing the fovea point theoretically required only 32.7ms. In our current implemented vergence control system using the available hardware (Intel Xeon CPU 3.2GHz, 2G Ram and Coreco PC2-Camlink frame grabbers), one cycle of image capturing and disparity estimation (both horizontal and vertical) cost about 200ms, resulting in a near real time control frequency.

Table 10 The processing time of the pyramidal model and the Coarse-to-Fine model

Image (size)	Processing time (seconds)	
	Pyramidal model	Coarse-to-Fine model [54]
Random dot (200×200)	7.18	68.34
Lab image (200×200)	10.81	129.35

5.3.4 Comparison with other related methods

The most relevant existing vergence control model in the literature is Sturzl's vergence control model using position-shift disparity energy neurons [47]. In Sturzl's model, a complex cell also contains a quadrature pair of simple cells and the disparity is encoded with the position shift of the receptive fields in the binocular images. The left receptive field of the complex cell was fixed at the reference fixation position and the right

receptive field of the complex cell shifted in a horizontal range to search for the position with highest energy. The position with highest energy indicates the match with the left receptive field. To handle large disparities, they used a discrete set of position shift with denser values near the fovea. For example, their groups of cells possessed position shifts of $[\pm 49, \pm 40, \pm 32, \pm 25, \pm 19, \pm 14, \pm 10, \pm 7, \pm 5, \pm 3, \pm 2, \pm 1, 0]$. The vergence control of Sturzl's method tries to locate the disparity channel which produces the minimum sum of energy in the whole image. Thus the vergence control attempts to verge towards an average depth in the scene instead of the depth of the fixation point. After vergence control, the fovea region may still possess a large amount of residual disparity.

Comparing with this method, our method has several advantages. Firstly, as the objective of vergence control should be minimizing the foveal disparity, the background disparities are inherently suppressed and the control focuses on the center. In the proposed method, this is achieved through the pyramidal approach where the effective receptive sizes were shrunk at each pyramidal level causing the disparity estimation become more localized. Secondly, Sturzl's method did not mention about vertical disparity that definitely exist in any robotic application environment. Thirdly, to handle large disparities, they used a discrete set of position shifts. Although the disparity range is larger, the same size of receptive field was used. Therefore there may be potential risk of missing target that is smaller than the disparity gap in the periphery. In contrast, the proposed model in this chapter also has some limitations. Firstly, the tuned disparity range is determined by the period of the Gabor filters used. As the pyramid progresses to each higher level to account for larger disparities, the image details are lost due to resolution change and this make the disparity estimation more difficult. Secondly, the computational load is heavier than the Sturzl's model given the same disparity range.

However, these shortcomings may be mitigated by avenues in parallel processing hardware that has become available.

For the purpose of comparison, Sturzl's vergence control model was adapted to the CogV system and tested in the same experimental environment. We used a discrete set of position shift and fixed cell size. Instead of minimizing the energy in the whole image, we minimize the energy at the center of the image. Otherwise the comparison is not reasonable because Sturzl's model always direct the control to an average depth of the scene. So the estimation is limited to the center of scene. The aspect ratio of the cells was set to 1. The controller is also a proportional controller and when the disparity is less than ± 1 , the control stops and vergence is achieved. Figure 63 shows the vergence control results with cell size 41 and searching range $[-40, 40]$. For most cases, the model worked correctly.

Figure 63 Vergence control results by Sturzl's method [47]: (a) vergence control results, (b) a wrong vergence and (c) a vertical misalignment (window cropped from image center)

However, for some positions, the model was misled by a higher peak that does not correspond to the correct position, as shown in Figure 63b. Figure 64 shows the initial status and the energy curve before vergence control for the situation in Figure 63b. There is a misleading peak in the right side of the curve, thus the system was directed to the wrong direction. A smaller searching range such as $[-20, 20]$ will resolve this problem. Another reason causing the vergence fail is the fixed window size. When there are sharp edges behind the foreground at different depth, the model tends to prefer the background depth as this is the common problem of all window based stereo matching approaches where the assumption is that the disparity in the matching window is uniform.

Figure 64 (a) The initial camera position before vergence control for the vergence result in Figure 63b, (b) the correct energy peak at disparity -7 was surpassed by another peak at disparity +40

To further compare the performance, five objects were put in front of the system and the proposed method and Sturzl's method were compared using the accuracy of depth estimation. The object displacement was shown in Figure 65.

Figure 65 The testing scene

Cell sizes and searching ranges for the Sturzl's method were varied. Generally speaking, our model performed better because it didn't fail for all five objects. Sturzl's method

failed several times for vergence on object 1 or 4. It again proved that a coarse-to-fine tuning may greatly improve the robustness of vergence control for cluttered scene.

Table 11 Depth estimation results

Object	The proposed pyramidal model	Sturzl's model		
		Cell size = 41 Search = -40:40	Cell size = 21 Search = -40:40	Cell size = 21 Search = -14:14
1	6.86% $\pm$ 1.54%	215.23% $\pm$ 5.06%	225.61% $\pm$ 8.94%	89.58% $\pm$ 18.16%
2	6.23% $\pm$ 3.18%	3.00% $\pm$ 1.23%	12.91% $\pm$ 2.60%	3.81% $\pm$ 2.05%
3	13.92% $\pm$ 4.96%	5.71% $\pm$ 7.87%	9.04% $\pm$ 5.62%	1.03% $\pm$ 1.68%
4	3.78% $\pm$ 3.37%	15.15% $\pm$ 5.07%	21.95% $\pm$ 3.19%	48.90% $\pm$ 2.24%
5	3.31% $\pm$ 4.22%	4.95% $\pm$ 1.82%	13.40% $\pm$ 0.73%	1.79% $\pm$ 2.47%

5.4 Discussions

The vergence control model presented in this chapter utilized a pyramidal approach coupled with the disparity energy model. As a result, there was a significant reduction of the computational load which provided the possibility for the system to function at practical vergence speeds for applications into active vision systems. To address the interference between vertical disparity and horizontal disparity, weighted compact orientation pooling was used to recover disparity and both horizontal and vertical disparities were minimized in vergence control. As an overall general conclusion, the proposed approach to the disparity energy model was used successfully in vergence control of a binocular pair of cameras in a natural indoor environment. It was compared to a previous position-shift disparity energy vergence control model and showed an overall better performance. With vergence capability, depth information of the fixation position can be directly derived using the motor configurations. The vergence control model is therefore suitable for near real time vergence control and can give a vivid set of depth information.

6 Type-C Vergence: A Pyramidal Cartesian Correlation Model for Vergence Control

Area correlation has been used widely in disparity estimation [66]. In area based stereo vision methods, the selection of window size is a common problem because the visual input has depth changes and the sizes of the target and the background region are unknown. Therefore coarse to fine algorithms have always been used to refine the disparity from a larger window to a smaller window. In this chapter, a vergence control model based on normalized cross correlation (NCC) between image patches in image pyramids is presented. This model is named PCC (Pyramidal Cartesian Correlation) model for convenience. In the PCC model, both the estimation of disparity and controlling of vergence are carried out sequentially through the layered pyramidal architecture. The pyramidal architecture provides a sequence of resolutions beginning with the coarser scale that provides for better coverage to the finer scale for accurate disparity estimations.

To incorporate color information, a pyramidal model with color masking was subsequently developed for vergence control. Instead of matching the whole image patch using NCC, an adaptive color mask was defined according to the color of the center pixel of the image patch. The matching between two image patches is through the average absolute difference between the corresponding pixels inside the color mask.

Therefore adaptive matching is achieved where the matching only focused on the visual information that is similar to the center of the image patch. The matching was also carried out in the image pyramidal architecture. This model is called the PCAD (Pyramidal Cartesian average Absolute Difference) model. The developed PCC and PCAD models were compared with a patch matching vergence control method [40] in the literature and showed robust performance.

6.1 The PCC vergence control model

The objective of vergence control system is to foveate the two cameras on a same unique point in the scene so that physical properties of the scene can be interpreted using the geometric principles first discovered by Charles Wheatstone [127]. Vergence difficulties arise when large disparity is presented in the fovea region. This can be overcome by performing disparity comparisons at coarser representations of the image. The PCC vergence control model leverages on a pyramidal image structure produced from the original binocular image pair. The disparity operator used was the normalized cross-correlation (NCC), which has proven to be effective in disparity estimation and tolerant to brightness changes in the image due to lighting and exposure conditions [25]. Disparity is recovered by searching for the position in the candidate image along a horizontal range having the maximum correlation with a position in the reference image. The NCC measure between two image patches is shown in Equation (48). The vergence control model targets only the localized region around the fixation point where the disparity comparisons mainly considers the center of the image, rather than the full image.

$$NCC = \frac{1}{n-1} \sum_{x,y} \frac{(f(x,y) - \bar{f})(g(x,y) - \bar{g})}{\sigma_f \sigma_g} \quad (48)$$

where

$$\sigma_f = \sqrt{\frac{1}{n-1} \sum_{x,y} (f(x,y) - \bar{f})^2} \quad (49)$$

$$\sigma_g = \sqrt{\frac{1}{n-1} \sum_{x,y} (g(x,y) - \bar{g})^2}$$

Equation (48) can be simplified to Equation (50), which can be considered as a measure for the similarity or fusion of the two patches.

$$NCC = \frac{\sum_{x,y} (f(x,y) - \bar{f})(g(x,y) - \bar{g})}{\sqrt{\sum_{x,y} (f(x,y) - \bar{f})^2 \sum_{x,y} (g(x,y) - \bar{g})^2}} \quad (50)$$

The original image is sub-sampled in a 2:1 ratio from lower to higher levels of the pyramid using Equation (51). The lowest level 0 (highest resolution) is the original image of size $M \times N$. In the vergence control system, a pyramid of 3 levels is used with resolutions of 200×200 , 100×100 and 50×50 respectively.

$$L_n(x, y) = L_0(x * 2^n, y * 2^n) \quad (51)$$

$$\text{where } x \in \left[0, \frac{M}{2^n} - 1\right] \text{ and } y \in \left[0, \frac{N}{2^n} - 1\right]$$

Figure 66 Image pyramid and disparity estimation

As shown in Figure 66, the disparity estimation process begins at the coarsest level and subsequently funnels through each level till the level of highest resolution is reached. Both horizontal and vertical disparities are estimated. At the coarsest level of $n-1$, the image patches at position 0 (image center) in the left image and position $[-d:d]$ in the right image along the mid-horizontal line are compared. Here $d=5$. The matching position is the position Pos_{n-1} with the highest NCC. The image patch has a size of $W \times W$. $W=11$ was used in the PCC model. The horizontal disparity for the center of the image is $H_{n-1} = right-left = Pos_{n-1}$. For vertical disparity, the estimation is after horizontal disparity estimation and is done at a column at the horizontal matching position Pos_{n-1} . The searching is in the vertical range $[-d_{n-1}:d_{n-1}]$. The position with maximum NCC is returned and denoted as $PosVer_{n-1}$ and the disparity $V_{n-1} = PosVer_{n-1}$. Hence a disparity pair (H_{n-1}, V_{n-1}) is obtained.

For the next level (level $n-2$), the search starts at the corresponding position of (H_{n-1}, V_{n-1}) at this level. The searching range with respect to this position is smaller and restricted to $[-d_{n-2}:d_{n-2}]$ where $d_{n-2}=1$, in both horizontal and vertical direction. The position $(Pos_{n-2}, PosVer_{n-2})$ with maximum NCC is selected. A disparity pair (H_{n-2}, V_{n-2}) is obtained. For the other levels, the estimation is the same as level $n-2$. This process continues until level 0 is arrived. The disparity pair (H_0, V_0) is the disparity for the image center.

The disparities for different levels of the pyramid can be used for vergence control in both horizontal and vertical direction. The controlling of slave camera is also coarse-to-fine and is an iterative process from level $n-1$ to level 0 until $H_0 = V_0 = 0$. This is similar to the vergence control process in section 5.2.4. In Figure 66, the camera is first shifted by (H_{n-1}, V_{n-1}) until the new (H_{n-1}, V_{n-1}) are less than a threshold. The control then proceeds to

the next level and repeats the same process. This continues until level 0 is reached. The control will stop if either the disparity (H_{n-1}, V_{n-1}) is within the threshold or oscillation occurs. Oscillation occurs when opposite disparities are encountered in consecutive vergence adjustments. Oscillations usually occur when there is insufficient variation in the fovea, or when occlusions are present in the image.

Apart from the physiological motivations for seeking a solution for vergence, a physical advantage is the imminent availability of a minimal disparity at the fovea when vergence is achieved. This property provides an inherent segmentation of the region by locating positions in the binocular image that possess zero disparity. It is possible then to distinguish the 3D binocular views of a door from the photograph of a real door. By applying a threshold on the disparity image of the image center, the region with minimal disparity ($\approx$ zero) can be extracted. Objects existing within this disparity range can be retrieved. The depth and 3D positions of the object in vergence can also be estimated.

6.2 Disparity estimation for vergence control

To illustrate the performance of the disparity estimation, the estimated disparity is plotted against the true disparity to support its effectiveness. The system was initially allowed to foveate on an object using the PCC vergence control model. The successful vergence is shown in Figure 67 where two cameras fixated on a single object, marked distinctly with a +, within in the scene.

Figure 67 Initial vergence: black crosses indicate the centers of the two images.

The right camera was subsequently manually panned from the fixation point over a range of $\pm 13^\circ$ in steps of 0.5° and the corresponding disparities were calculated. The resultant disparity is plotted in Figure 68a.

Figure 68 (a) The disparity estimation curve and (b) the curve superimposed with the left image

The results indicate that the disparity estimation curve between -5 and 5 is a relatively straight line bearing a slope of approximately 45° . This implies that the center part of the curve is ideal for vergence control. For illustrative clarity, the disparity curve is superimposed on the actual camera image in Figure 68b. The 0 mark in the intersection of the axes coincides with the foveated position of the cameras. Note also that at the $\pm 7^\circ$ position, the disparity drops to 0. These correspond to the positions of the mug and

remote control and are local peaks. Therefore it is possible to attain local correlation peaks when the right camera fixates on an out-of-view object that is in close visual proximity. This is a drawback of the limited search range and fixed window size, but can be better resolved if visual accommodation factors in the visual system are added to improve the initial saccade allocation [9, 55, 139].

6.3 Experimental results

6.3.1 Combining saccade and vergence: exploration of the scene

The system was allowed to perform a series of autonomous saccades using the FLANN segmentation model with local information checking presented in section 3.5.2. Figure 69 shows the resultant sequence of the center of images. The left and right images are shown in pairs and the system was successful in generating saccades with good vergence on the interesting areas in the scene. It is also noted that the binocular image columns in Figure 69 can be perceived in 3D by the human observer. The fused image will emerge from the page as a 3D perceptive image. This is an unquantifiable indication of successful vergence.

Figure 69 Exploration of the scene: the fovea images in the process of saccade and vergence.

6.3.2 The zero-disparity region

The zero-disparity regions of the fovea of two cameras can be easily retrieved by setting a threshold on disparity. Here a threshold of ± 1 is used. Figure 70a shows the zero-disparity regions retrieved. The regions that possess higher disparity values are blot out in grey. The disparity estimation is carried out in the area of 40×40 in the 200×200 resolution image pyramid. The zero-disparity mask generated was enlarged by a ratio of 5 and put on the fovea of the finer image of 1000×1000 resolution. Through the process of NCC comparisons, many tiny regions were generated in the segmentation. To eliminate these tiny regions, three consecutive binary morphological closing operations [97] were applied on the zero-disparity masks. The second row of Figure 70b shows the results after the morphological operation. The mask is now less noisy and can be used to retrieve the zero disparity regions effectively. Note the regions are bigger than the objects because the uniform background also exhibits zero disparity. The segmentation is important to the learning of objects and can improve the performance of recognition because redundant information beyond the fused region can be removed. This aids the figure-ground separation process.

Figure 70 Segmentation of objects based on disparity: (a) regions based solely on disparity and (b) regions after morphological operations.

6.3.3 Depth estimation

A depth estimation experiment was conducted to evaluate the vergence performance. Seven objects (Figure 67) were put in front of the CogV system and the system sequentially located and autonomously verged on the objects. Upon vergence, the pan-tilt configurations of the system were used to calculate the object depth. The system was executed three times and the mean of the calculations were taken. The average error of estimation and the standard deviation of error are shown in Table 12. One percent error corresponds to about 2cm in depth at a depth of 2m. The errors obtained were less than 3% (about 6cm). Considering that an object had a body width of about 5-10 cm, the estimation of depth was considerably accurate.

Table 12 Depth estimation results

Object	Average Estimated Depth (m)	True Depth (m)	Average Estimation Error
Box	2.11	2.16	2.51%±0.50%
Rod	2.01	2.07	2.57%±0.00%
Bottle	1.50	1.52	1.15%±0.00%
Mug	1.81	1.89	4.46%±0.31%
Trishaw	1.73	1.75	1.31%±1.07%
Remote Control	1.61	1.62	0.60%±0.00%
Green Bottle	1.37	1.38	0.57%±0.32%

6.4 The PCAD vergence control model: extension to color images with color masking

Color images provide more fruitful information than grayscale images. In this section, a color masking based disparity estimation method is used in the pyramidal architecture for vergence control. The NCC measure is replaced with the average absolute difference (AAD). The matching local window is covered by a mask generated by the differences

between each pixel and the center pixel. This is based on the assumption that pixels with similar color tend to belong to the same object and hence is from the same depth. This reduces the contribution from other depth regions not consistent with the foreground. This is similar to segmentation based stereo methods where the image is segmented and disparity in an image segment is assumed to be consistent [140].

RGB images are used for the disparity estimation. To compensate noises in the image capturing process, the average color of a 3×3 neighborhood at the image center is selected as the reference color O . The pixels in the matching window are compared with O . The binary color mask is generated by Equation (52). If a pixel in the matching window has a difference smaller than a threshold T in all 3 color dimensions, this position is set to 1 in the mask. In the current implementation, T is set to 30 for the color image with a [0,255] range in each dimension.

$$M(x, y) = \begin{cases} 1, & \text{if } |I_c(x, y) - O_c| < T \text{ for } c \in [R, G, B] \\ 0, & \text{else} \end{cases} \quad (52)$$

Figure 71 shows the generated masks for several sample image patches. Through the masking process, background contribution to the similarity measure is minimized. Although there may be some contribution from different depth sharing similar color property with the center pixel, these regions are typically smaller than the foreground and contribute less to the similarity measure. By applying the mask to both left and right images, the matching is limited to the region that is most relevant to the foreground of the left image.

Figure 71 The image patches and the generated masks

To further address the situation that the foreground may be textured or consists of small segments, the mask is enlarged if the number of pixels with value 1 in the mask is too small. This is by setting a region at the center of mask to 1 if the number of pixels with value 1 in the mask occupies less than 10% of the matching window. The region is defined by a square having half of the image patch length and width. In this case, the matching becomes a normal window based matching approach rather than using a distorted mask.

Average absolute difference (AAD) in the *RGB* domain is used as the dissimilarity measure between image patches. Disparity estimation is based on a shifting window approach while the AAD measure is only applied to those pixels with value 1 in the color mask. Due to the variable sizes of color mask, AAD was used instead of SAD (sum of absolute difference). A pyramidal window based approach is also adopted here. The architecture of the algorithm is similar to the PCC model.

Figure 72 (a) Vergence control with the PCC model and (b) vergence control with the PCAD model

Objects with different hue but the same luminance level can be distinguished. Figure 72 shows the result of vergence for the PCC model and the PCAD model. Initially both cameras are parallel and the left camera fixated on the brown mug. The PCC model produced a higher peak for the blue mug and directed the right camera to the blue mug. Note that in Figure 72a the brown mug in the left image and the blue mug in the right image shared a similar local background. This is why the right camera was directed to the blue mug because in grayscale images, both the foreground and background possess a higher degree of matching. However, the PCAD model was not confused by the blue mug and its background. It directed the right camera to the correct position because in color images, the mask provides the foreground region and the AAD provides a more accurate dissimilarity measure. If the color masking is applied to the left image of Figure 72a, the mask generated is shown in Figure 73. The contribution to the AAD measure

mainly comes from the brown mug. Therefore the brown cup in the right image will have a higher AAD than the blue mug.

Figure 73 An image and its color mask

It is also expected that the masking process gives more accurate disparity estimation because information from different disparity or depth region can be largely removed due to different color profile with the reference color.

6.5 Comparison with existing methods

The relevant existing vergence control models in the literature include those that used local area correlation for patch matching, such as Yamato's method where LoG (Laplacian of Gaussian) filtered images and SAD (Sum of Absolute Distances) were used. The disparity was determined by the position in the right image which has minimum SAD with the reference position in the left image. For comparison purpose, Yamato's method was adapted into the CogV system. The reference position is the center of the left image and the search range is $[-30, 30]$ along the horizontal scan line with respect to the center of the right image. The same depth estimation experiment as that in Section 5.3.4 was conducted. The estimation result was shown in Table 13. Yamato's method with window size 21 showed an overall best result while the result of the proposed method is also comparable. Yamato's method with window size 41 failed on object 4 where the

estimation was misled to a wrong SAD extrema. Therefore for general vergence control, the selection of window size is a critical issue. Comparing to Yamato's method, the pyramidal approach provides a natural way of resolving the out of range disparity and reducing matching confusion in a coarse-to-fine manner. However, potential problem of the proposed method would be a misleading peak at coarser level which may project to the finer level and cause estimation error.

Table 13 Depth estimation results

Object	PCC	PCAD	Yamato's	
			Window size = 41	Window size = 21
1	3.81±0.34%	3.79±1.25%	5.08±1.82%	2.98±1.00%
2	1.13±0.01%	0.98±0.55%	0.88±0.01%	1.11±0.39%
3	2.67±2.33%	1.64±1.80%	0.09±0.03%	0.64±0.59%
4	0.68±0.58%	0.89±0.84%	194.44±166.78%	1.20±0.38%
5	3.90±0.96%	4.40±2.30%	3.29±2.48%	2.69±1.92%

6.6 Discussions

In this chapter two pyramidal models for real time vergence control were presented. The experimental results showed that the developed system was able to autonomously explore a simple visual environment by performing successful vergence on the points of interest. Positional information of the objects in the environment can be extracted and derived. This allows the efficient collection of information and provides useful information for vision based localization and navigation tasks. The method was also extended to color images by an adaptive color masking approach. Better foreground matching can be achieved with less confusion. The PCC and PCAD models were also compared with an existing area correlation method for vergence control and showed more robust performance.

7 Type-D Vergence: A Pyramidal Log Polar Correlation Model for Vergence Control

The visual-cortical mapping in the human vision system can be modeled by the log polar transformation [19]. This transformation magnifies the center of the visual field, providing high acuity for the point of fixation. Vergence control requires the accurate estimation of the disparity at the center of image and the central magnification property of log polar transformation provides good ways of concentrating on the matching of the center of images. Previous chapters have also shown that coarse to fine method through image pyramids reduces the computation while preserving the performance. Based on these, this chapter provides a study on the suitability of log polar images for vergence control and presented a robust and stable vergence control model combining the pyramidal approach and log polar transformation. The stability of disparity estimation in multiple-resolution log polar images made it possible to develop a pyramidal vergence control method. The binocular images were converted into two image pyramids and subjected to the coarse-to-fine disparity estimation using normalized cross correlation over the log polar space. The proposed model is named the PLP (Pyramidal Log Polar correlation) model. With the PLP model, the CogV system was able to verge on objects in a complex environment in real time. The system was also shown to function successfully even under different focusing and brightness of the binocular images, rendering it suitable for real world robot vision applications. The model was compared with a

previous log polar transformation based vergence control model [59] and showed more stable and robust performance.

7.1 PLP: the pyramidal log polar correlation model

One major problem in vergence control is the significant depth changes between foreground and background. Techniques that assuming a smooth depth surface in a local area may suffer from this and require supporting techniques such as foreground segmentation, avoiding the information from background. In this section this problem is addressed by the use of log polar images. Taking advantage of the cortical magnification, the contribution from the image center has a much larger weight than the periphery. This provides a robust way of estimating the fovea disparity for vergence control.

7.1.1 Normalized cross correlation in log polar domain

Normalized cross correlation (NCC) is invariant to camera gain and offset variation [67] and has often been used in disparity estimation and vergence control [39, 141]. NCC defines the cross correlation between two normalized image patches, shown in Equation (53). In Chapter 0 it has been shown to be stable and effective in area based stereo matching in the Cartesian image domain. Since the human visual cortex is known to be structured in a log polar design and the cortical magnification property of log polar transform has the potential to provide better vergence, we explored the path of combining both the NCC with the log polar based cortical map.

$$NCC = \frac{\sum_{x,y} (f(x,y) - \bar{f})(g(x,y) - \bar{g})}{\sqrt{\sum_{x,y} (f(x,y) - \bar{f})^2 \sum_{x,y} (g(x,y) - \bar{g})^2}} \quad (53)$$

Corresponding patches in log polar domain cannot be directly matched. A linear shift in Cartesian domain transforms to a non-linear shift in log polar domain. The correlation

based method becomes more complex in log polar domain because the matching window also gets non-linearly transformed. However, no matter what the shift is and how the matching window gets transformed, the matching position still preserves the highest correlation with the reference position [83]. It is also expected that only when the two images coincide, especially the image centers coincide, the two images have highest correlation in log polar domain.

Figure 74 The binocular image pair

A simple experiment was conducted to evaluate the effectiveness of using log polar image for disparity estimation. Figure 74 shows a pair of binocular images. The foveas of the left and right image were fixated at the same point in space, targeting the bottle. To derive a disparity tuning curve the left camera was fixed while the right camera was systematically panned to the left and right at fixed steps (about 0.5°). A series of binocular image pairs were generated. These images were used to evaluate the suitability of disparity estimation using log polar images. Comparisons with shifting window correlation based disparity estimation were also carried out. Figure 75 shows the comparison between matching two images in log polar space and Cartesian space.

Figure 75 Disparity tuning curves with log polar and Cartesian images: (a) varying log polar image resolution, (b) varying Cartesian image and transform to fixed size log polar images and (c) varying window sizes in Cartesian domain.

Figure 75a shows the relationship between the log polar resolution and the disparity estimation. At each position, the images were transformed to log polar image of a specified resolution and the NCC between the left log polar image and the right log polar image was calculated. NCC values were plotted against the true disparity to provide a disparity tuning curve. The disparity tuning curves for different resolutions of log polar images showed no obvious difference indicating that the estimation is highly stable regardless the resolution of log polar images. The center peak of the disparity tuning curve indicates that NCC in log polar domain is effective in determining the vergence status and is applicable for vergence control. Figure 75b shows the disparity tuning curves by fixing the log polar resolution and varying the original binocular image resolution. The curve shows stability and the reduction of resolution in Cartesian space is insignificant to the correlation value. This is important for the approach developed in this chapter because a pyramidal image structure was utilized and the invariance to resolution provides good performance in the coarser levels of the pyramid. In comparison Figure 75c shows the window based Cartesian space disparity estimation. Large variations occur for different window sizes. The results indicate a high reliance on the windowed spatial information and it was difficult to define a proper matching

window because the foreground size is unknown. Through this experiment, the log polar space is shown to be a significantly more robust space for vergence control. The bio-inspired log polar images can be considered to provide a robust adaptive matching window by varying the contributions of each pixel according to its proximity to the center of the image.

7.1.2 An evaluation of disparity measures and image space

This section provides an evaluation of different disparity measures and image space. SAD (sum of absolute difference), SSD (sum of squared difference), NCC, and NCC in log polar were compared. Zero-mean methods are invariant to intensity offset. Previous research has shown that among these correlation measures only NCC is invariant to both camera gain and intensity offset [67]. Images captured from the CogV system were used for the evaluation to show the advantages of log polar images. Figure 76 shows the three reference images. The foreground was put in front of the system at different distance so that in the master camera it occupies 50%, 25% and 12.5% of the full image width. The slave camera was panned to the left and right in a wide range at a step of 2.5° . Images were captured from the slave camera and the matching between the reference image and the collected series of slave images were calculated and plotted against the vergence angle to produce a disparity tuning curve.

Figure 76 The left image of the experiments. The foreground object width is (a) 50% (b) 25% and (c) 12.5% of the full image width.

The results using different similarity measures were shown in Figure 77. All the curves were normalized by their maximum and (1-NCC) was used instead of NCC for comparison purpose. In Figure 77, the background (the wall and posters) vergence angle and the foreground (the paper lid) vergence angle were plotted as dotted and solid vertical lines respectively. It can be seen that as a global measure, the SAD, SSD and NCC were all biased by the background disparity. In this case a proper matching window should be selected to focus on the foreground, which requires a foreground segmentation process. The NCC using log polar images was however not biased and shows correct tuning curve with the correct position having a minimum value.

Figure 77 Disparity tuning curves for SAD, SSD, NCC and NCC in log polar space. The solid vertical line refers to the foreground, and the dotted vertical line refers to the wall and door behind.

7.1.3 Robust vergence in log polar domain

As log polar image is a cortical magnified stretching of the original image, the correlation in log polar image is a type of weighted correlation in the Cartesian space. This weightage can be derived as the magnification ratio of log polar transform where its ratio is defined as the number of pixels in log polar space that corresponds to one pixel at a certain location in Cartesian space. This ratio can be decomposed into two components: the magnification along ρ axis and along φ axis. In a continuous transformation, one pixel in Cartesian domain at a radial distance r occupies a step from $r-0.5$ to $r+0.5$ along ρ axis. This interval corresponds to the following distance on axis ρ :

$$R_\rho = \log_a(r + 0.5) - \log_a(r - 0.5) = \frac{d(\log_a r)}{dr} = \frac{1}{r \ln(a)} \quad (54)$$

Along the φ axis, one ring of pixels with diameter r corresponds to φ_{max} pixels in log polar space. Therefore the magnification ratio on φ axis is:

$$R_\varphi = \frac{\varphi_{max}}{2\pi r} \quad (55)$$

The overall magnification ratio is the multiplication of these two components.

$$R = R_\rho R_\varphi = \frac{\varphi_{max}}{2\pi r^2 \ln(a)} = \frac{\varphi_{max}}{2\pi \ln(a)} \frac{1}{(x - x_0)^2 + (y - y_0)^2} \quad (56)$$

The magnification ratio is therefore inverse proportional to the squared distance from the origin of the log polar transformation. In correlation based disparity estimation in log polar space, the contribution of a pixel to the correlation measure will be determined by this magnification ratio. Assuming the system is fixating on a foreground object, the correlation measure can be considered as summations of the correlation from foreground region and background region. The image can be decomposed into two components.

$$I = I_{fore} + I_{back}$$

Note that this foreground-background definition is based on the reference image, such as the left image in the CogV system. For a binocular image pair I_L and I_R , the correlation measure can be considered as summations of the correlation from foreground region and background region separately.

$$\begin{aligned} C(I_L, I_R) &= C_{fore}(I_L, I_R) + C_{back}(I_L, I_R) \\ &= C(I_{fore}^L, I_{fore}^R) + C(I_{back}^L, I_{back}^R) \end{aligned}$$

Considering the log polar transformation as a weight matrix W applied to the original Cartesian image to change the weight of each pixel wise correlation, W can also be decomposed to two components W_{fore} and W_{back} . The correlation in log polar domain can be expressed as:

$$\begin{aligned} C(P_L, P_R) &= C_{fore}(P_L, P_R) + C_{back}(P_L, P_R) \\ &\equiv C(I_{fore}^L * W_{fore}, I_{fore}^R * W_{fore}) + C(I_{back}^L * W_{back}, I_{back}^R * W_{back}) \end{aligned}$$

where P_L and P_R are the log polar images

The weight for the foreground region is significantly larger than the background as the foreground is at the fovea of the input image for vergence control tasks. When the foreground regions of the binocular image pair match and produce a high correlation, the changes of background contributes little to the correlation measure because of the exponential decreasing of magnification ratio. The background is attenuated. However, when the foreground regions do not fuse, both the foreground correlation and background correlation are small. As an overall result, the disparity tuning curve has a distinct peak at the correct disparity. This significantly stabilizes the vergence on foreground object, especially when the foreground is small and has large disparity changes with the background. In reverse, in Cartesian domain, due to the even contribution from the foreground and background, vergence is directed to an

intermediate depth, depending on the sizes of foreground and background regions in the matching window.

7.1.4 Coarse to fine disparity estimation using image pyramid

Coarse to fine strategy uses multi-level processing to enlarge the working range of the disparity estimation and at the same time reduce computation [54]. The coarser resolution result is used as an initial estimation for the finer level. In the PLP model, a binocular image pyramid is used to estimate disparity in a coarse to fine manner. The original binocular image pair is sub-sampled to form a binocular image pyramid through Equation (57). The images are sub-sampled in a 2:1 ratio between consecutive levels of the pyramid. The lowest level (highest resolution) is level 0 which is the original image of size $M \times N$ and the resolution for level n in the pyramid is $M/2^n \times N/2^n$.

$$L_n(x, y) = L_0(x * 2^n, y * 2^n) \quad (57)$$

$$\text{where } x \in \left[0, \frac{M}{2^n} - 1\right] \text{ and } y \in \left[0, \frac{N}{2^n} - 1\right]$$

In the PLP model a pyramid of 3 levels, level 0 to 2, was used, as shown in Figure 78. The resolutions of each level are 200×200 , 100×100 and 50×50 respectively. The disparity estimation process goes from the coarsest resolution to the finest resolution.

Figure 78 The log polar transformation in the image pyramid

The log base a of the log polar transform is determined by the top level (coarsest) image size (M_T, N_T) and the transformed log polar image size $(\rho_{max}, \varphi_{max})$ according to Equation (58). The transformed log polar image covers a maximum inner circle of the top level image. The log base a is calculated so that the top level image of size 50×50 is converted to a 50×90 log polar image, using log polar sampling of the original image. For the lower levels, this log base a is fixed and covers the same size of area in pixels. In Figure 78, due to the fixed log base a , the log polar transformation covers a region that is becoming smaller downward the Cartesian domain image pyramid. This shrinking of regions down the pyramid makes the algorithm focus more and more on the local information. The magnification of the fovea region increases down the pyramid. In this way, foreground disparities contribute more to the final disparity estimation.

$$a = \exp\left(\frac{\ln\left(\min\left(\frac{M_T}{2}, \frac{N_T}{2}\right)\right)}{\rho_{max}}\right) \quad (58)$$

Taking the left image as the reference image and the right image as the candidate image, the disparity estimation process is based on the shift of the origins of log polar transformation in the candidate image. For a pixel (x_L, y_L) in the left image, the objective is to find the corresponding pixel $(x_L + d_H, y_L + d_V)$ in the right image, where the (d_H, d_V) pair is the horizontal and vertical disparity. The left image is transformed to log polar image with (x_L, y_L) as the origin. For the right image, the origin for log polar transformation is shifted horizontally with respect to the reference position. We get a series of log polar images with a series of shifted origins, shown in Figure 79 (b). The right log polar image with the maximum NCC with the left log polar image determines the disparity at this position. When the origin is near corner or boundary of the image, the corresponding

positions of some of the log polar pixels may be out of the original image. In this case, the values of these log polar pixels are simply set to 0.

Figure 79 Disparity estimation using log polar image: black pixel denotes the reference position in the left image and the candidate positions in the right image. (a) Transformation of the left image with the reference position as origin and (b) transformation of the right image with the candidate positions as origins

Considering the log polar transformation in Equation (59):

$$\begin{aligned}\rho &= \log_a \sqrt{(x - x_0)^2 + (y - y_0)^2} \\ \varphi &= \arctan\left(\frac{y - y_0}{x - x_0}\right)\end{aligned}\tag{59}$$

The log polar sampling is actually an inverse transform of Equation (59) and is shown in Equation (60). The mapping can be pre-calculated to increase the processing speed. The shifting of origin can be realized by simply adding an offset d to x in the mapping matrix.

$$\begin{aligned}x &= \begin{cases} x_0 + \sqrt{\frac{a^2 \rho}{1 + \tan^2(\varphi)}}, & \varphi \in \left[0, \frac{\pi}{2}\right] \text{ or } \left[\frac{3\pi}{2}, 2\pi\right] \\ x_0 - \sqrt{\frac{a^2 \rho}{1 + \tan^2(\varphi)}}, & \text{otherwise} \end{cases} \\ y &= y_0 + x \tan(\varphi)\end{aligned}\tag{60}$$

The coarse to fine process is realized through disparity inheritance between successive levels. A disparity range for searching is set at each level. The algorithm starts from the top level. The disparity at this level is sent to the next level. In the next level, the starting reference position in the right image is shifted by the disparity so that the searching

starts nearer to the true corresponding position. The estimation in the lower level is restricted in a narrower range compared to the top level processing. Let n be the number of levels in the pyramid. In the pyramid we have level 0 to level $n-1$, from finest to coarsest resolution. The disparity estimation is firstly carried out at the coarsest level, level $n-1$. The searching radius was set to 10 for level $n-1$ and 3 for the rest levels. According to Equation (61), the 3-level pyramid model can cover a maximum disparity of ± 49 pixels, which is half the visual field of the 200×200 image. The computation time is only $(10+3+3)/49=33\%$ compared to using a direct searching method covering the same range of disparity.

$$\text{Disparity range } D = \sum_{l=0}^{n-1} d_l * 2^l \quad (61)$$

$$\text{Computational time } T = \sum_{l=0}^{n-1} d_l$$

where d_l is the searching radius for level l

For the vertical disparity, the estimation at each level starts after the horizontal disparity estimation. The starting position in the right image is shifted according to the estimated horizontal disparity and disparity estimation is then carried out in a vertical range. The position with maximum NCC determines the vertical disparity. The searching range is a constant range $[-3, 3]$ for each level because the binocular system has a horizontal baseline which has smaller disparity in vertical direction. However, vertical disparity affects the estimation of horizontal disparity and must be corrected.

The advantage of using log polar images in a pyramidal approach is that matching is performed at the global image but systematically reduces to a localized estimation of disparity. The PLP model takes advantage of both the log polar transformation and the pyramidal image structures to stabilize vergence on foreground and enlarge the working

range at low cost. It provides both the speed and robustness, ensuring an accurate match in the final level.

The control of vergence adopts a similar strategy to those presented in section 5.2.3. The left camera was assigned as the master camera and the right camera as the slave camera. The controlling of slave eye is an iterative process from the coarsest level to finest level. The controller gain was set to 0.8 for robust and stable vergence. Oscillation was detected in the control process.

7.2 Experimental results

Both qualitative and quantitative experimental results are shown in this section. The first set of experiments shows the vergence control results. The system was tested in a cluttered lab environment and was shown to verge successfully on objects in the scene. In the second set of experiments, the system was also shown to be invariant to focusing and brightness changes of the lens. The third set of experiments shows depth estimation results, which indirectly verifies the vergence control model. Comparisons between the PLP model and the previously developed models were presented. Vergence control speeds of the available models were also compared. The fourth set of experiments shows the combination of saccade and vergence to provide a general reconstruction of depth information of important regions in the scene.

7.2.1 Vergence control

Vergence control was tested in a lab environment with a wide range of disparities in the scene. Three objects were put in front of the cameras at a distance about 1 meter and the background lab environment is about 2-7 meters away. A series of binocular image

pairs after vergence control were captured and presented in Figure 80. The system successfully verged on the foreground objects, the background objects and the ceiling. The system was able to achieve successful stable final state for all runs of vergence control with no oscillation. The system showed robust performance when depth changed significantly in the fovea area. For example, this can be seen from the first image pair in the second row of Figure 80. The foreground has a bottle at the fovea. The background however was quite different for the two views. The plants and other objects through the door have significant disparity changes. The system was still able to verge on the bottle taking the advantage of the cortical magnification of log polar transformation. For the same situation, if we apply the PCC model which used the NCC in Cartesian space instead of log polar space, the model became unstable (large oscillation happened) and was not able to verge on the bottle. This is the advantage of the log polar transformation over the normal Cartesian space method.

Figure 80 Vergence control results

The right camera was subsequently adjusted to a different level of brightness. The system still showed promising performance. The results are shown in Figure 81. This is very important for real world robotic application where the hardware may suffer more disturbances from the environment and should be able to adapt to different conditions.

Figure 81 Vergence control results when the two cameras were set to different brightness

The vergence control was also tested in a corridor scene and the results are shown in Figure 82. The proposed model also showed very good performance in this scene. For robot applications such as indoor navigation, this scene is a common environment and the vergence control model will help in determining the distance from the platform to the target under fixation.

Figure 82 Vergence control results in a corridor scene

Figure 83 shows the situation when a foreground object is removed in the middle of vergence control. The system initially verged on the foreground, which is the hand. In this case, if the pyramidal correlation method in Cartesian space was used, the slave camera did not stop on the hand but was directed to the background instead. This is because the background has large disparity difference with the foreground, and the foreground occupies a small area at the fovea of the image. A uniform matching window received more contribution from the background and hence was not able to verge on the foreground. The log polar based method however was able to verge correctly

because of the fovea magnification. After the first vergence on the hand, the foreground object was removed and the system continues to adjust its vergence status and successfully verged on the background at a further depth. In this experiment, the proportional controller gain is set to 0.2 to make the vergence control slower for the capturing of image sequences during vergence.

Figure 83 Vergence control when the target is removed: (a) binocular images at the beginning of vergence control and (b) fovea region sequence under vergence control.

The vergence angles for the image sequence in Figure 83 were shown in Figure 84a. Time step 11 is the moment when the hand is removed from the fovea. The fixation depth was also derived and plotted against the time step, shown in Figure 84b. The time step 11 corresponds to a depth of about 1.3 meters, which is approximately the true depth of the hand.

Figure 84 Vergence control when the target is removed: (a) vergence angles and (b) estimated depth

7.2.2 Tolerance to focusing and brightness

In this section, the tolerance of the PLP model to focusing and brightness changes was systematically tested by manually setting one camera of the system to a different focusing or brightness level. Figure 85 shows the captured images under different focusing conditions. The left camera was set to focus on a distance from far to near, creating the left image 1 to 7 in Figure 85. Vergence on an object at depth 1.80m was used to evaluate the performance.

Figure 85 The left camera was adjusted to 7 levels of focusing

The depth of the target object was calculated based on the vergence results and used to evaluate the performance under different focusing conditions. Table 14 shows the depth

estimation error. Under the seven levels of focusing, the system was able to successfully verge on the object and estimate the depth accurately (accuracy > 96%).

Table 14 Depth estimation under different focusing

Focusing setting	Estimation error
1	3.26%
2	3.36%
3	1.51%
4	3.26%
5	1.56%
6	3.19%
7	1.51%

Subsequently the two cameras were set to the same focusing setting but the left camera was set to different levels of brightness by changing the aperture of the lens. Figure 86 shows the captured images under the different brightness.

Figure 86 The left camera was adjusted to 7 levels of brightness

The result of depth estimation is shown in Table 15. The system successfully verged on the object for the first six setting of brightness. The seventh setting exceeded the

threshold brightness and most areas in the image became white causing the system to fail. The estimated depth was irrelevant and gave a very large error.

Table 15 Depth estimation under different brightness

Brightness level	Estimation error
1	3.14%
2	3.36%
3	4.00%
4	2.69%
5	3.19%
6	3.41%
7	160.47%

Through this experiment, the PLP model was shown to have robust performance under different alterations of the hardware. This makes it suitable for real world applications.

7.2.3 Depth estimation

Depth estimation was also used to evaluate the performance of vergence in a lab scene with various settings of foreground and background. The PLP model was compared with the other available methods presented in the previous chapters. Objects were put in front of the system and the system was directed to the objects for vergence control. Once vergence control is completed, the pan tilt configurations of the two cameras were used to calculate the depth of the target using straightforward geometrical calculation. Figure 87a shows a view of the experimental setup and the objects are marked with '+'. The detailed placement of objects is shown in Figure 87b. Vergence control were carried out for the five objects in a sequential manner. When the two cameras are parallel, the initial disparity for the object at 1.8m depth is about 30 pixels in the 200×200 image.

Figure 87 The experimental environment: (a) a camera view and (b) the placement of objects. Numbers near the object indicates the height of the object

The depth estimation result of the PLP is shown in Table 16. The PLP model showed promising performance on depth estimation with average accuracy above 90% and very small standard deviation of error (<1%). For both the nearer and further objects, the model shows very accurate depth estimation.

Table 16 Depth estimation results of the PLP model

Object	True depth (m)	Average estimated depth (m)	Average absolute error	Std dev of error
1	1.80	1.86	3.42%	0.48%
2	1.80	1.79	0.66%	0.40%
3	2.80	2.83	0.79%	0.63%
4	2.30	2.29	1.19%	1.20%
5	2.80	2.71	3.06%	0.21%

For the purpose of comparison, the depth estimation was also conducted using the developed vergence control models in the previous chapters. The Type-A log polar edge method in chapter 4 was not able to perform successful vergence in this cluttered environment because the log polar edges generated didn't come from the same set of objects. Therefore this model is omitted from comparison. The results that follow

compare the Type-B pyramidal disparity energy model in Chapter 0 and the two Type-C pyramidal Cartesian models in Chapter 0 with the Type-D log polar correlation model. The color masking was also used in the Type-D PLP model with AAD as the disparity operator. The mask is generated by comparing the difference between the pixels in log polar domain and the center reference pixel in Cartesian domain. If the generated mask is less than 10% of the log polar image, the region with $\rho < \frac{1}{3}\rho_{max}$ of the log polar image was set to 1 in the mask. This model is called PLP-CM (PLP with Color Masking) model. The models for comparisons are summarized in Table 17.

Table 17 Models in the depth estimation experiment

Model	Description
Type-B	Pyramidal disparity energy model
Type-C1	PCC: the pyramidal Cartesian correlation model
Type-C2	PCAD: the pyramidal AAD model with color masking
Type-D1	PLP: the pyramidal log polar model
Type-D2	PLPCM: the pyramidal log polar model with color masking

The method that is most relevant to the PLP model in the literature is Bernadino's model [59] which also uses log polar images and NCC for vergence control. The control strategy is similar to Sturzl's method. The disparity range is a discrete inhomogeneous set of disparities. This method was adapted to the CogV system with a range of disparity $[\pm 49, \pm 40, \pm 32, \pm 25, \pm 19, \pm 14, \pm 10, \pm 7, \pm 5, \pm 3, \pm 2, \pm 1, 0]$. This range is equivalent to the model presented in this chapter (the maximum disparity is ± 49). For each disparity, an inverse log polar warp was defined for the candidate image and the correlation between the reference log polar image and the warped candidate log polar image was used for disparity estimation. In the following section, Sturzl's model, Yamato's model and Bernadino's model were tested for comparison with the proposed models. Depth

estimation was carried out three times for each model. Table 18 shows the average absolute error and standard deviation of error of the models.

Table 18 Depth estimation error and standard deviation of error

Object	Type-B	Type-C1	Type-C2	Type-D1	Type-D2
1	6.86±1.54%	3.81±0.34%	3.79±1.25%	3.42±0.48%	3.87±1.06%
2	6.23±3.18%	1.13±0.01%	0.98±0.55%	0.66±0.40%	1.78±0.38%
3	13.92±4.96%	2.67±2.33%	1.64±1.80%	0.79±0.63%	3.44±1.95%
4	3.78±3.37%	0.68±0.58%	0.89±0.84%	1.19±1.20%	0.68±0.34%
5	3.31±4.22%	3.90±0.96%	4.40±2.30%	3.06±0.21%	3.17±0.94%

Object	Yamato's Window = 21	Sturzl's Cell size = 41 Search = -40:40	Bernadino's Search = -49:49
1	2.98±1.00%	215.23±5.06%	3.77±1.69%
2	1.11±0.39%	3.00±1.23%	9.93±7.47%
3	0.64±0.59%	5.71±7.87%	2.55±1.07%
4	1.20±0.38%	15.15±5.07%	2.59±2.44%
5	2.69±1.92%	4.95±1.82%	167.36±284.96%

The overall results showed that among the developed models Type-D1 model has the best performance. Type-C1, Type-C2 and Type-D2 models also showed robust performance compared to Type-B model. Type-B disparity energy model shows poorer performance compared to other developed models. This is because in cluttered scenes, depth changes significantly in a local area. The assumption of the disparity energy model is that the two input visual stimuli are horizontally shifted version of each other. This may not be true in the experiment, especially at the coarser level of the pyramid.

Yamato's method with window size 21 showed good performance. Sturzl's method with cell size 41 showed poorer performance for object 1 and 4. These two objects have relatively large depth gap with their background and their background contains more

distractive vertical lines compared to other objects' local background. Bernadino's method showed a general good performance but failed for object 5. Compared to other object, this object is relatively small thus Bernadino's method may not be able to derive the disparity because it has a larger step at periphery disparity range. If the object falls in the gap, it won't give a significant peak in the disparity tuning curve. The Type-D1 and Type-D2 model succeeded in this case due to the foveation downward the image pyramid and the searching is pixel wise so no visual information is stepped over in the disparity tuning curve. The fovea region is also magnified more and more downward the pyramid.

7.2.4 Real time performance

The computational load of the available models was also compared. The systems were allowed to performance vergence on several objects in the lab scene. The average time taken to complete one vergence control cycle was shown in Table 19. The Type-B pyramidal disparity model provided a near real time performance with control frequency about 5Hz. All other developed models took less than 30ms to process one cycle, including the video memory accessing and disparity computation. This leads to control frequency above 30Hz which fulfills the real time requirement because the video frame rate of the cameras is only 20Hz. Sturzl's method showed very fast performance. This is because the implementation of Sturzl's method in the CogV system is just to retrieve corresponding patches from the video buffer (41×41) and the filtering process is a straightforward matrix multiplication. Yamato's model was a bit slower due to the LoG filtering process of the captured image (200×200 image filtered by 9×9 kernels). Bernadino's method showed a faster speed than the PLP model because it has only one level processing without having the need to construct the image pyramid.

Table 19 Processing time of the models

Model	Setting	One cycle processing time
Type-B	Cell size=29, 3 levels	205.8ms
Type-C1	Window size=11×11, 3 levels	8.24ms
Type-C2	Window size=11×11, 3 levels	8.96ms
Type-D1	Image size=50×90, 3 levels	28.35ms
Type-D2	Image size=50×90, 3 levels	24.85ms
Sturzl's model	Cell size=41, disparity=-40:40	<1.00ms
Yamato's model	Window size=21, disparity=-30:30	44.89ms
Yamato's model	Window size=41, disparity=-30:30	47.30ms
Bernadino's model	Image size=50×90, disparity=-49:49	6.18ms

7.2.5 Combining saccade and vergence: exploration of the scene

The vergence model was then incorporated with the FLANN segmentation and color attention model as saccade generator so that the system can function autonomously. The autonomous saccadic vergence results are presented. The motor positions were recorded down, transformed to the image domain and superimposed on the image captured at the beginning of the experiment. Figure 88a,d shows the saccade generated. Figure 88b,e shows the vergence movements of the slave camera. The generated saccades showed good correspondence with informative positions in the scene.

Figure 88 (a, d) Saccadic movement by the master camera, (b, e) saccadic and vergence movements by the slave camera and (c, f) Reconstruction of the positions of fixated objects. Cameras were shown as triangles and objects were shown as circles.

The positions of the objects under fixation are an important source of information for navigation and other positional operations such as object manipulation. Figure 88c,f shows the 3D position reconstructed. It can be seen that the reconstructed positions of the fixation positions indicate good consistency with the true object positions. The depth estimation capability can be extended to various applications such as using vision to do a horizontal sweep of vergence and reconstruct the open space in front of a robot. The robot can use this information for obstacle avoidance or path planning. Another example is that when a mobile robot is moving forward and the binocular vision system on the robot is facing front, the vergence status of the two cameras directly gives the depth information in the moving direction. The robot can use this information to make decisions early and avoid collision.

7.2.6 Comparison with existing log polar vergence control model

Comparing with existing log polar vergence control methods, our method has several advantages. For those methods using log polar sensors [46], direct correlation matching is not possible because of the non-linear transformation. Geometrical transforms between the log polar space and the Cartesian space for each disparity value need to be calculated separately for the correspondence definition. Based on the idea that the correct vergence position gives the highest correlation value and is a local peak in the disparity tuning curve, one way of the vergence control is a derivative based control which climbs to the peak of the disparity tuning curve [59]. Apparently this is very sensitive to variations of the disparity tuning curve and may be easily trapped to local minimum. Bernadino's method used an inverse log polar warp to match the image. This is not as efficient as the backward sampling used in the PLP model. The PLP model pre-computes the log polar transformation with respect to origin and a shift is simply added to the whole transformation matrix to generate the backward log polar mapping matrix. We also showed that the disparity estimation in log polar space has tolerance to the resolution of log polar images and the input Cartesian images. Therefore this pyramidal log polar correlation model was developed to efficiently compute the disparity within the image. The pyramidal approach made it possible to cover a dense range of disparities and provided robust performance.

Every algorithm has its working range. Although the log polar transform magnifies the image center, the target object still has a lower limit of the size. This is the size where the foreground object occupies a smaller region in the log polar image than the background region. In the experiment conducted in Figure 76, we have shown that the

system can accommodate an object that is as small as 12.5% of the view field in Cartesian space. Another problem of vergence is the confusing visual information as shown in Figure 60, the combination of foreground and background happens to provide a better confusing matching position. This problem is however not solvable by disparity estimation method because if the visual stimulus produces better matching at wrong disparities, the disparity estimation method cannot differentiate the wrong-position good matching and correct-position good matching. This problem can be solved by considering other components of the human vision system. For example, accommodation is a direct method to provide this capability. If there are multiple high peaks in the disparity tuning curve, the peak nearest to the depth indicated by best focusing of the lens should be the correct position for fixation. Due to the limitation of hardware, this issue cannot be implemented in the CogV system but the cooperation of accommodation and disparity based vergence is certainly an important future research topic for active vision systems. Existing studies have largely relied on the accommodation to provide the initial status for vergence [39, 55] which is a practical engineering approach while the human vision system does not assume this privilege of accommodation. As a whole system, multiple components of the vergence mechanism and their interactions need to be investigated and simulated.

7.3 Discussions

A spatial variant approach for vergence control is presented in this chapter. This model has the following advantages. Firstly, the model was inspired by the log polar transformations existing in the visual-cortical pathway of the primate vision system. The assumption of vergence control is that the fixation of master eye is on a foreground target of unknown size. The fovea magnification property of the log polar transformation

was utilized for accurate and reliable vergence control. Through the experiments, stable and accurate vergence was achieved in a scene with a wide range of disparity. The developed model was shown to be more invariant to disparity gaps and background changes than the normal Cartesian space methods. This is important for real world robot applications where reliability and flexibility are important. Secondly, the log polar transformation also reduces the computational load as a relatively low resolution log polar image can provide the robust and stable vergence performance. The model was deployed in the CogV system with real time performance. As a general conclusion, the model is suitable for real world robot vision applications.

8 Conclusions and Future Work

This thesis provides a focused study on the simulation of vergence eye movements in active vision systems. From the studies on biological vision systems, bio-inspired computational vergence models were better understood and developed. An active vision platform was built to explore visual scene using saccade and vergence. The characteristics of the human vision system gave the system inspirations in both theoretical design and platform implementation.

8.1 Summary

The saccadic eye movements were simulated using image segmentation techniques. A watershed network was initially developed for segmentation by detecting the edge features and simulating pool filling. This provided the basic object segmentation in a restricted scene with a plain background. After this, a Fast Learning Artificial Neural Network (FLANN) color segmentation algorithm was developed to segment the input image to blobs and generate saccade according to local feature and information around the segment centers. Color attention model was also combined with the FLANN segmentation model for target selection. These models served as information selection mechanisms in the developed CogV system. Important visual information could be retrieved by the system.

Several bio-inspired vergence models were developed. Related difficulties and problems were investigated and solved through computational methods. Vergence control can be derived from different mechanisms such as disparity estimation and accommodation. This thesis focuses on disparity based vergence control. Vergence control models are expected to have wide working range, tolerance to noise and low computational cost because the primary objective is for real world applications and the application environment may not be as ideal as those used for dense stereo map reconstruction.

Four types of vergence control models were developed. The initial log polar model simulates the visual cortical mapping, center-surround processing and visual pathway in the early vision system. This model works well in a constrained environment but fails to verge in cluttered scenes. Subsequently based on the study on binocular visual neurons, a pyramidal approach to disparity energy model was developed and was able to provide vergence in a cluttered lab environment. Near real time performance was achieved using the developed model. After this, a pyramidal Cartesian correlation model was developed to provide adequate performance. Eventually the pyramidal log polar correlation model was developed to provide both stable and robust vergence, especially in cluttered environment. The cortical magnification of the visual-cortical transform was utilized for high concentration on the foreground object. This provides a fast and robust vergence control model applicable for real world applications.

Saccade and vergence control models are basic components of an artificial cognitive vision system. With these capabilities, the system is able to capture visual information effectively. This is helpful for higher level visual recognition and visual navigation systems. For an autonomous robot, the capability of actively determining the location

and exploring the local spatial information are important. With the groundwork developed in this thesis, a robot with binocular vision is now a viable option that is able to robustly provide depth for exploration of scenes. This provides a new avenue for active binocular systems to achieve a step closer for cognitive visual behavior. Interactions between robots and humans can now be complemented with eye-contact and visual systems can reduce their processing scopes to a selected area of attention instead of processing an entire image.

8.2 Possible future work

In this thesis saccade generation is carried out through image segmentation techniques. The visual attention components simulated the bottom up attention. However, in real world applications, task dependant information is important. Therefore task dependant saccade generation is an interesting topic for future research. A viable way to construct task dependant saccade generation is through the incorporation of top down attention [142].

As the human naturally possesses a spatial variant retina, the spatial variant log polar images should be naturally related to the human visual processing in the brain. The saccade generation, especially the visual attention mechanism, may be adapted to log polar images. Other components such as accommodation and motion detection can be explored for the suitability in the log polar space. With the reduced resolution in periphery, the log polar image may be suitable for peripheral motion detection which can be used for the generation of saccadic eye movements.

The human vision faculties are combined efforts of the sensory, motor and different cortex areas. A simple vision task often requires the cooperation of several visual mechanisms. It is thus reasonable to develop multimodal system to address a single visual function. As the vergence control in human vision is not solely based on disparity, possible future work in vergence is the development of multimodal vergence control models. The multimodal cooperation can be realized through including other components of vergence such as accommodation or through including other types of visual information such as motion and proximal cues for vergence. Interaction between components of vergence is important. Some existing systems incorporate motion or accommodation as the selection of initial saccade [39]. In our investigation, the geometry of vergence control was analyzed and disparity was used as the sole source for vergence. Inclusion of other components will definitely improve the vergence performance. Besides this, various applications of vergence control in robot systems can also be explored. Possible applications would be those requiring accurate positioning capability such as grasping, obstacle avoidance or surface modeling.

References

- [1] D. H. Hubel, *Eye, brain, and vision*. New York: W. H. Freeman, 1988.
- [2] C. Busettini and L. E. Mays, "Saccade–vergence interactions in macaques. I. test of the omnipause multiply model," *Journal of Neurophysiology*, vol. 94, pp. 2295–2311, 2005.
- [3] F. Pardo, B. Dierickx, and D. Scheffer, "CMOS foveated image sensor: signal scaling and small geometry effects," *IEEE Transactions on Electron Devices*, vol. 44, pp. 1731-1737, 1997.
- [4] R. Wodnicki, G. W. Roberts, and M. D. Levine, "A log-polar image sensor fabricated in a standard 1.2- μm ASIC CMOS process," *IEEE Journal of Solid-state Circuits*, vol. 32, pp. 1274-1277, 1997.
- [5] C. Capurro, F. Panerai, and G. Sandini, "Dynamic vergence using log-polar images," *International Journal of Computer Vision*, vol. 24, pp. 79-94, 1997.
- [6] M. M. Marefat, L. Wu, and C. C. Yang, "Gaze stabilization in active vision - I. vergence error extraction," *Pattern Recognition*, vol. 30, pp. 1829-1842, 1997.
- [7] S. Grossberg and A. Yazdanbakhsh, "Laminar cortical dynamics of 3D surface perception: stratification, transparency, and neon color spreading," *Vision Research*, vol. 45, pp. 1725-1743, 2005.
- [8] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Deigo, CA, USA, 2005, pp. 995-1000.
- [9] L. R. Squire, F. E. Bloom, S. K. McConnell, J. L. Roberts, N. C. Spitzer, and M. J. Zigmond, *Fundamental neuroscience*, 2 ed. San Diego, California, USA: Academic Press, 2003.
- [10] E. Hecht, *Optics*, 4 ed. New York: Addison Wesley, 2002.
- [11] D. H. Hubel and T. N. Weisel, "Sequence regularity and geometry of orientation columns in the monkey striate cortex," *Journal of Computational Neurology*, vol. 158, pp. 267-293, 1974.
- [12] Y. Yeshurun and E. L. Schwartz, "Cortical hypercolumn size determines stereo fusion limits," *Biological Cybernetics*, vol. 80, pp. 117-129, 1991.
- [13] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106-154, 1962.
- [14] R. B. Tootell, M. S. Silverman, E. Switkes, and R. L. D. Valios, "Deoxyglucose analysis of retinotopic organization in primate striate cortex," *Science*, vol. 218, pp. 902-904, 1982.

- [15] P. M. Daniel and D. Whitteridge, "The representation of the visual field on the cerebral cortex in monkeys," *Journal of Physiology*, vol. 159, pp. 203-221, 1961.
- [16] B. Fischer, "Overlap of receptive field centers and representation of the visual field in the cat's optic tract," *Vision Research*, vol. 13, pp. 2113-2120, 1973.
- [17] D. C. Van Essen, W. T. Newsome, and J. H. Maunsell, "The visual representation in striate cortex of macaque monkey: asymmetries, anisotropies, and individual variability," *Vision Research*, vol. 24, pp. 429-448, 1984.
- [18] E. L. Schwartz, "Afferent geometry in the primate visual cortex and the generation of neuronal trigger features," *Biological Cybernetics*, vol. 28, pp. 1-24, 1977.
- [19] E. L. Schwartz, "Spatial mapping in the primate sensory projection: analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, pp. 181-194, 1977.
- [20] J. Ross, D. Burr, and C. Morrone, "Suppression of the magnocellular pathway during saccades," *Behavioural Brain Research*, vol. 80, pp. 1-8, 1996.
- [21] M. Mon-Williams and H. C. Dijkerman, "The use of vergence information in the programming of prehension," *Experimental Brain Research*, vol. 128, pp. 578-582, 1999.
- [22] D. R. Melmoth and S. Grant, "Advantages of binocular vision for the control of reaching and grasping," *Experimental Brain Research*, vol. 171, pp. 371-388, 2006.
- [23] A. L. Yarbus, *Eye movements and vision*. New York: Plenum Press, 1967.
- [24] D. A. Robinson, "Eye movements evoked by collicular stimulation in the alert monkey," *Vision Research*, vol. 12, pp. 1795-1808, 1972.
- [25] T. J. Olson, "Stereopsis for Verging Systems," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1993, pp. 55-60.
- [26] K.-O. Ludwig, H. Neumann, and B. Neumann, "Local stereoscopic depth estimation," *Image and Vision Computing*, vol. 12, pp. 16-34, 1994.
- [27] J. J. Clark, "Spatial attention and latencies of saccadic eye movements," *Vision Research*, vol. 39, pp. 585-602, 1999.
- [28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, Nov 1998.
- [29] L. Itti, "Models of bottom-up and top-down visual attention," *PhD Thesis, California Institute of Technology*, 2000.
- [30] K. Brunnstrom, J. O. Eklundh, and T. Uhlin, "Active fixation for scene exploration," *International Journal of Computer Vision*, vol. 17, pp. 137-162, February 1996.

- [31] G. J. Giefing, H. Jan, and H. Mallot, "Saccadic object recognition with an active vision system," in *International Conference on Pattern Recognition*, The Hague, Netherlands, 1992, pp. 664-667.
- [32] S. Balasuriya and P. Siebert, "An architecture for object-based saccade generation using a biologically inspired self-organised retina," in *International Joint Conference on Neural Networks*, Vancouver, BC, Canada, 2006, pp. 4255-4261.
- [33] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.
- [34] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature*, vol. 391, pp. 481-484, 1998.
- [35] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," in *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, USA, 1994, pp. 222-228.
- [36] M. Khokhotva, H. Ono, and A. P. Mapp, "The cyclopean eye is relevant for predicting visual direction," *Vision Research*, vol. 45, pp. 2339-2345, 2005.
- [37] J. P. Siebert and D. F. Wilson, "Foveated vergence and stereo," in *International Conference on Visual Search*, Nottingham, UK, 1992.
- [38] J. Taylor, T. Olson, and W. N. Martin, "Accurate Vergence Control in Complex Scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 540-545.
- [39] W.-S. Ching, P.-S. Toh, and M.-H. Er, "Robust vergence with concurrent detection of occlusion and specular highlights," *Computer Vision and Image Understanding*, vol. 62, pp. 298-308, 1995.
- [40] J. Yamato, "A layered control system for stereo vision head with vergence," in *IEEE International Conference on Systems, Man and Cybernetics*, Tokyo, Japan, 1999, pp. 836 - 84.
- [41] X. Roca, J. Vitria, M. Vanrell, and J. J. Villanueva, "Gaze control in a binocular robot systems," in *IEEE International Conference on Emerging Technologies and Factory Automation*, Barcelona, Spain, 1999.
- [42] N. Ahuja and A. L. Abbott, "Active stereo: integrating disparity, vergence, focus, aperture, and calibration for surface estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1007-1029, 1993.
- [43] W. M. Theimer and H. A. Mallot, "Phase-based binocular vergence control and depth reconstruction using active vision," *CVGIP: Image Understanding*, vol. 60, pp. 343-358, 1994.
- [44] M. Hansen and G. Sommer, "Real-time vergence control using local phase differences," *Machine Graphics and Vision*, vol. 5, pp. 51-63, 1996.

- [45] Y. Yeshurun and E. L. Schwartz, "Cepstral Filtering on a Columnar Image Architecture: A Fast Algorithm for Binocular Stereo Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 759-767, July 1989.
- [46] R. Manzotti, A. Gasteratos, G. Metta, and G. Sandini, "Disparity estimation on log-polar images and vergence control," *Computer Vision and Image Understanding*, vol. 83, pp. 97-117, 2001.
- [47] W. Sturzl, U. Hoffmann, and H. A. Mallot, "Vergence control and disparity estimation with energy neurons: theory and implementation," in *International Conference on Artificial Neural Networks*, Madrid, Spain, 2002, pp. 1255-1260.
- [48] A. Gibaldi, M. Chessa, A. Canessa, F. Solari, and S. P. Sabatini, "A neural model for binocular vergence control without explicit calculation of disparity," in *European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2009.
- [49] E. K. C. Tsang, S. Y. M. Lam, Y. Meng, and B. E. Shi, "Neuromorphic implementation of active gaze and vergence control," in *IEEE International Symposium on Circuits and Systems*, Seattle, WA, 2008, pp. 1076 - 1079.
- [50] E. Krotkov, K. Henriksen, and R. Kories, "Stereo ranging with verging cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 1200-1205, 1990.
- [51] P. Peixoto, J. Batista, and H. Araujo, "A surveillance system combining peripheral and foveated motion tracking," in *International Conference on Pattern Recognition*, 1998.
- [52] C. Yim and A. C. Bovik, "Vergence control using a hierarchical image structure," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, 1994, pp. 118 - 123.
- [53] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 204, pp. 301-328, 1979.
- [54] Y. Chen and N. Qian, "A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms," *Neural Computation*, vol. 16, pp. 1545-1577, 2004.
- [55] A. L. Abbott and N. Ahuja, "Surface reconstruction by dynamic integration of focus, camera vergence, and stereo," in *International Conference on Computer Vision*, Bombay, India, 1988.
- [56] M. Hansen and G. Sommer, "Active depth estimation with gaze and vergence control using Gabor filters," in *International Conference on Pattern Recognition*, Vienna, Austria, 1996, pp. 287 - 291.
- [57] T. J. Olson and D. J. Coombs, "Real-time vergence control for binocular robots," *International Journal of Computer Vision*, vol. 7, pp. 67-89, 1991.

- [58] N. Oshiro, N. Maru, A. Nishikawa, and F. Miyazaki, "Binocular tracking using log polar mapping," in *International Conference on Intelligent Robots and Systems*, 1996, pp. 791-798.
- [59] A. Bernardino and J. Santos-Victor, "Vergence control for robotic heads using log-polar images," in *International Conference on Intelligent Robots and Systems*, Japan, 1996.
- [60] J. H. Piater, R. A. Grupen, and K. Ramamritham, "Learning real-time stereo vergence control," in *International Symposium on Intelligent Control/Intelligent Systems and Semiotics*, Cambridge, MA, USA, 1999, pp. 272-277.
- [61] S.-B. Choi, B.-S. Jung, S.-W. Ban, H. Niitsuma, and M. Lee, "Biologically motivated vergence control system using human-like selective attention model," *Neurocomputing*, vol. 69, pp. 537-558, 2006.
- [62] J. Batista and P. Peixoto, "Real-time vergence and binocular gaze control," in *International Conference on Intelligent Robots and Systems*, 1997.
- [63] J. Batista, P. Peixoto, and H. Araujo, "Binocular tracking and accommodation controlled by retinal motion flow," in *International Conference on Pattern Recognition*, Barcelona, Spain, 2000.
- [64] H.-J. Kim, M.-H. Yoo, and S.-W. Lee, "A control model for vergence movement on a stereo robotic head using disparity flux," in *International Conference on Pattern Recognition*, Barcelona, Spain, 2000.
- [65] R. Marfil, C. Urdiales, J. A. Rodríguez, and F. Sandoval, "Automatic vergence control based on hierarchical segmentation of stereo pairs," *International Journal of Imaging Systems and Technology*, vol. 13, pp. 224-233, 2003.
- [66] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7-42, 2002.
- [67] J. Banks and P. Corke, "Quantitative evaluation of matching methods and validity measures for stereo vision," *The International Journal of Robotics Research*, vol. 20, pp. 512-532, 2001.
- [68] K.-J. Yoon and S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 650-656, 2006.
- [69] D. J. Fleet and A. D. Jepson, "Stability of phase information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1253-1268, 1993.
- [70] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understanding*, vol. 53, pp. 198-210, 1991.
- [71] T. D. Sanger, "Stereo disparity computation using Gabor filters," *Biological Cybernetics*, vol. 59, pp. 405-418, 1988.

- [72] W. M. Theimer, H. A. Mallot, and S. Tolg, "Phase method for binocular vergence control and depth reconstruction," *Proceedings of SPIE*, vol. 1826, pp. 76-87, 1992.
- [73] N. P. McLoughlin and S. Grossberg, "Cortical computation of stereo disparity," *Vision Research*, vol. 38, pp. 91-99, 1998.
- [74] E. K. C. Tsang and B. E. Shi, "Normalization enables robust validation of disparity estimates from neural populations," *Neural Computation*, vol. 20, pp. 2464-2490, 2008.
- [75] G. F. Poggio and B. Fischer, "Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey," *Journal of Neurophysiology*, vol. 40, pp. 1392-1405, 1977.
- [76] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Depth is encoded in the visual cortex by a specialized receptive field structure," *Nature*, vol. 352, pp. 156-159, 1991.
- [77] D. J. Fleet, H. Wagner, and D. J. Heeger, "Neural encoding of binocular disparity: energy models, position shifts and phase shifts," *Vision Research*, vol. 36, pp. 1839-1857, 1996.
- [78] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors," *Science*, vol. 249, pp. 1037-1041, 1990.
- [79] N. Qian and Y. Zhu, "Physiological computation of binocular disparity," *Vision Research*, vol. 37, pp. 1811-1827, 1997.
- [80] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2D visual cortical filters," *Journal of Optical Society America. A*, vol. 2, pp. 1160-1169, 1985.
- [81] B. P. Bogert, W. J. R. Healy, and J. W. Tukey, "The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Symposium on Time Series Analysis*, New York, 1963, pp. 209-243.
- [82] D. J. Coombs and C. M. Brown., "Intelligent gaze control in binocular vision," in *IEEE International Symposium on Intelligent Control*, Philadelphia, PA, USA, 1990, pp. 239-245.
- [83] E. Grosso, R. Manzotti, R. Tiso, and G. Sandini, "A space-variant approach to oculomotor control," in *International Symposium on Computer Vision*, Coral Gables , Florida , USA, 1995, p. 509.
- [84] E. Grosso, R. Manzotti, R. Tiso, and G. Sandini, "A space-variant approach to oculomotor control," in *The International Symposium on Computer Vision*, 1995, p. 509.

- [85] A. Bernardino and J. Santos-Victor, "Binocular visual tracking: Integration of perception and control," *IEEE Transactions on Robotics Automation*, vol. 15, pp. 1080-1094, 1999.
- [86] L. E. Mays, "Neural control of vergence eye movements," in *The Visual Neurosciences*. vol. 2, L. M. Chalupa and J. S. Werner, Eds., ed: MIT Press, 2004, pp. 1415-1426.
- [87] J. G. Samarawickrama and S. P. Sabatini, "Version and vergence control of a stereo camera head by fitting the movement into the Hering's Law," in *Canadian Conference on Computer and Robot Vision*, 2007.
- [88] E. Hering, *The Theory of Binocular Vision*. New York: Plenum Press, 1977.
- [89] H. Collewijn, C. J. Erkelens, and R. M. Steinman, "Trajectories of the human binocular fixation point during conjugate and non-conjugate gaze-shifts," *Vision Research*, vol. 37, pp. 1049-1069, 1997.
- [90] J. T. Enright, "Monocularly programmed human saccades during vergence changes?," *Journal of Physiology*, vol. 512, pp. 235-250, 1998.
- [91] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, pp. 54-115, 1987.
- [92] A. L. P. Tay and S. Prakash, "K-Means fast learning artificial neural network, an alternative network for classification," in *International Conference on Neural Information Processing*, Singapore, 2002, pp. 925-929.
- [93] L. P. Tay, J. M. Zurada, L.-P. Wong, and J. Xu, "The hierarchical fast learning artificial neural network (HieFLANN) - an autonomous platform for hierarchical neural network construction," *IEEE Transactions on Neural Networks*, vol. 18, pp. 1645-1657, Nov 2008.
- [94] E. M. Klier, H. Wang, and J. D. Grawford, "The superior colliculus encodes gaze commands in retinal coordinates," *Nature Neuroscience*, vol. 4, pp. 627-632, 2001.
- [95] S. Buecher and C. Lantuejoul, "Use of watersheds in contour detection," in *International Workshop on Image Processing*, Rennes, France, 1979, pp. 2.1-2.12.
- [96] X. Zhang and A. L. P. Tay, "Fast learning artificial neural network (FLANN) based color image segmentation in R-G-B-S-V cluster space," in *International Joint Conference on Neural Networks*, Orlando, Florida, USA, 2007, pp. 563-568.
- [97] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 2nd ed. New Jersey: Prentice Hall, 2002.
- [98] J. Horvath, "Image segmentation using clustering," in *Proceedings of 3rd Doctoral Conference*, Košice, 2003, pp. 105-106.

- [99] M. Niemeier, J. D. Crawford, and D. B. Tweed, "Optimal transsaccadic integration explains distorted spatial perception," *Nature*, vol. 422, pp. 76-80, 2003.
- [100] S. Marcelja, "Mathematical description of the responses of simple cortical cells," *Journal of Optical Society America*, vol. 70, pp. 1297-1300, 1980.
- [101] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of Optical Society America. A*, vol. 4, pp. 2379-2394, 1987.
- [102] R. Methrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern Recognition*, vol. 25, pp. 1479-1494, 1992.
- [103] X. Gao, F. Sattar, and R. Venkateswarlu, "Corner detection of gray level images using Gabor wavelets," in *International Conference on Image Processing*, Singapore, 2004, pp. 2669-2672.
- [104] N. Ludtke, B. Luo, E. Hancock, and R. C. Wilson, "Corner detection using a mixture model of edge orientation," in *International Conference on Pattern Recognition*, Quebec City, Canada, 2002, pp. 574-577.
- [105] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE Transactions on Image Processing*, vol. 12, pp. 729-739, 2003.
- [106] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 1-13, 1996.
- [107] C. Liu and H. Wechsler, "Independent component analysis of Gabor features for face recognition," *IEEE Transactions on Image Processing*, vol. 14, pp. 919-927, 2003.
- [108] D. Dunn and W. E. Higgins, "Optimal Gabor filters for texture segmentation," *IEEE Transactions on Image Processing*, vol. 4, pp. 947-964, 1995.
- [109] H. Wu, Y. Yoshida, and T. Shioyama, "Optimal Gabor filters for high speed face identification," in *International Conference on Pattern Recognition*, Québec City, Canada, 2002.
- [110] L. Chen, G. Lu, and D. Zhang, "Effects of different Gabor filter parameters on image retrieval by texture," in *International Multimedia Modelling Conference*, Brisbane, Australia, 2004.
- [111] S.-Y. Cho and J. J. Wong, "Robust facial recognition by localized Gabor features," in *International Workshop on Advance Imaging Technology*, Korea, 2005.
- [112] D. Dunn, W. E. Higgins, and J. Wakeley, "Texture segmentation using 2-D Gabor elementary functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 130-149, 1994.

- [113] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Transactions on Image Processing*, vol. 11, pp. 1160-1167, 2002.
- [114] J.-K. Kamarainen, V. Kyrki, and H. Kälviäinen, "Invariance properties of Gabor filter-based features—overview and applications," *IEEE Transactions on Image Processing*, vol. 15, pp. 1088-1099, 2006.
- [115] P. Kruizinga and N. Petkov, "Nonlinear operator for oriented texture," *IEEE Transactions on Image Processing*, vol. 8, pp. 1395-1407, 1999.
- [116] C.-J. Lee and S.-D. Wang, "A Gabor filter-based approach to fingerprint recognition," in *IEEE Workshop on Signal Processing Systems*, 1999, pp. 371-378.
- [117] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 837-842, 1996.
- [118] T. N. Tan, "Rotation invariant texture features and their use in automatic script identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 751-756, 1998.
- [119] J. P. Jones and A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, pp. 1233–1258, 1987.
- [120] The USC-SIPI Image Database [Online]. Available: <http://sipi.usc.edu/database/>
- [121] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Data Processing*. vol. 1, D. E. Rumelhart and J. McClelland, Eds., ed Cambridge, MA: The M.I.T. Press, 1986, pp. 318-362.
- [122] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.
- [123] A. Yonas, *Perceptual development in infancy*. Hillsdale, New Jersey Hove and London: Lawrence Erlbaum Associates Inc, 1988.
- [124] B. G. Cumming, "An unexpected specialization for horizontal disparity in primate primary visual cortex," *Nature*, vol. 418, pp. 633-636, 2002.
- [125] J. P. Siebert and I. Eising, "Scale-space recognition based on the retino-cortical transform," in *IEE Conference on Image Processing and its Applications*, Edinburgh, 1995.
- [126] R. A. Peters, M. Bishay, and T. Rogers, "On the computation of the log-polar transform," ed: Technical Report, Intelligent Robotics Laboratory, Center for Intelligent Systems, Vanderbilt University School of Engineering, 1996.

- [127] C. Wheatstone, "Contributions to the physiology of vision. - part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision," *Philosophical Transactions of the Royal Society of London*, pp. 371-394, 1838.
- [128] N. Qian, "Binocular disparity and the perception of depth," *Neuron*, vol. 18, pp. 359-368, 1997.
- [129] D. Gabor, "Theory of communication," *Journal the Institution of Electrical Engineers*, vol. 93, pp. 429-459, 1946.
- [130] H. Zhang, B. Zhang, W. Huang, and Q. Tian, "Gabor wavelet associative memory for face recognition," *IEEE Transactions on Neural Networks*, vol. 16, pp. 275-278, 2005.
- [131] B. S. Manjunath and R. Chellappa, "A unified approach to boundary perception: edges, textures, and illusory contours," *IEEE Transactions on Neural Networks*, vol. 4, pp. 96-108, 1993.
- [132] C. Chang and S. Chatterjee, "Ranging through gabor logons - a consistent, hierarchical approach," *IEEE Transactions on Neural Networks*, vol. 4, pp. 827-843, 1993.
- [133] A. Arleo, F. Smeraldi, and W. Gerstner, "Cognitive navigation based on nonuniform Gabor space sampling, unsupervised growing networks, and reinforcement learning," *IEEE Transactions on Neural Networks*, vol. 15, pp. 639-652, 2004.
- [134] N. Petkov, "Biologically motivated computationally intensive approaches to image pattern recognition," *Future Generation Computer Systems*, vol. 11, pp. 451-465, 1995.
- [135] P. O. Bishop and J. D. Pettigrew, "Neural mechanisms of binocular vision," *Vision Research*, vol. 26, pp. 1587-1600, 1986.
- [136] A. Nieder and H. Wagner, "Encoding of both vertical and horizontal disparity in random-dot stereograms by wulst neurons of awake barn owls," *Visual neuroscience*, vol. 18, pp. 541-547, 2001.
- [137] U. R. Dhond and J. K. Aggarwal, "Structure from stereo - a review," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, pp. 1489-1510, 1989.
- [138] J. T. Enright, "Monocularly programmed human saccades during vergence changes," *Journal of Physiology*, vol. 512, pp. 235-250, October 1998.
- [139] W.-S. Ching, P.-S. Toh, K.-L. Chan, and M.-H. Er, "Robust vergence with concurrent detection of occlusion and specular highlights," in *International Conference on Computer Vision*, Berlin, Germany, 1993.
- [140] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *International Conference on Image Processing*, Singapore, 2004, pp. 2997- 3000.

- [141] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 920-932, 1994.
- [142] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, 2007.

Index

Active vision systems	33	Ocular dominance columns	24, 99
		Orientation pooling	125
Central visual pathway	99		
Cepstrum	48	Panum's fusional area	32
CogV	57, 58	Phase-shift model	118
Color masking	155	Position-shift model	118
Correlation measures	43	Pyramidal Cartesian correlation model	147
		Pyramidal disparity energy model	114
Disparity energy model	46, 115	Pyramidal log polar correlation model	161
Eye	21	Receptive field	25
Eye movements	28	Retina	21
		Retino-cortical map	27
FLANN	53		
FLANN segmentation model	65	Saccade integration	75
		Spatial pooling	127
Gabor filters	78		
Gabor signatures	84	Tonic vergence	92
Gaze shifting	29		
Gaze stabilization	29	Vergence control	38
		Vergence control systems	42
		Vergence geometry	90
Human vision system	21	Vergence in log polar domain	167
		Version	51
Inverse log polar transform	171	Visual attention model	35
		Visual pathway	22
KFLANN	55		
Log polar transformation	27, 96	Watershed model	62

Appendices

Appendix A. The derivation of phase shift disparity, in association with the work of Chen and Qian [54]

Assume the left image is $I(x, y)$ and the right image is $I(x + D, y)$ having a disparity D with the left image. Then the right image is a left-shifted version of the left image and D is the difference between the horizontal coordinate of a point in the left image and its corresponding point in the right image. The simple cell response can be written as

$$r_s = \iint_{-\infty}^{\infty} \{g_l(x, y)I(x, y) + g_r(x, y)I(x + D, y)\}dxdy$$

Since shifting the left image by D is equivalent to shifting the filter by $-D$, we can write

$$r_s = \iint_{-\infty}^{\infty} \{g_l(x, y)I(x, y) + g_r(x - D, y)I(x, y)\}dxdy$$

In the above equation, $g_l(x, y)$ can be written as:

$$\begin{aligned} g_l(x, y) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega x + \varphi) \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) [\cos(\omega x) \cos(\varphi) - \sin(\omega x) \sin(\varphi)] \\ &= \cos(\varphi) g_{\cos}(x, y) - \sin(\varphi) g_{\sin}(x, y) \end{aligned}$$

$$\text{where } g_{\cos}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{\sigma_y^2}\right) \cos(\omega x)$$

$$\text{and } g_{\sin}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{\sigma_y^2}\right) \sin(\omega x)$$

Separate the left and right receptive field, Let

$$r_s = r_{sl} + r_{sr} = \iint_{-\infty}^{\infty} g_l(x, y)I(x, y)dxdy + \iint_{-\infty}^{\infty} g_r(x - D, y)I(x, y)dxdy$$

Then $r_{sl} = \iint_{-\infty}^{\infty} g_l(x, y)I(x, y)dxdy$

$$= \iint_{-\infty}^{\infty} [\cos(\varphi) g_{\cos}(x, y) - \sin(\varphi) g_{\sin}(x, y)]I(x, y)dxdy$$

$$= \cos(\varphi) \iint_{-\infty}^{\infty} g_{\cos}(x, y)I(x, y)dxdy - \sin(\varphi) \iint_{-\infty}^{\infty} g_{\sin}(x, y)I(x, y) dxdy$$

$$= A \cos(\alpha + \varphi)$$

$$\text{where } A = \sqrt{A_1^2 + A_2^2} \text{ and } \alpha = \arctan\left(\frac{A_2}{A_1}\right)$$

$$\text{and } A_1 = \iint_{-\infty}^{\infty} g_{\cos}(x, y) I(x, y) dx dy \text{ and } A_2 = \iint_{-\infty}^{\infty} g_{\sin}(x, y) I(x, y) dx dy$$

For the right receptive field and the phase shift model:

$$g_r^{pha}(x - D, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-D)^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega(x - D) + \varphi + \Delta\varphi)$$

The next is the first order Taylor expansion in $\frac{D}{\sigma_x}$ for the term $\exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right)$ in $g_r^{pha}(x - D, y)$. The Taylor expansion of a signal $f(x)$ is:

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \dots$$

So we have $f\left(\frac{D}{\sigma_x}\right) = \exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right)$ and $f'\left(\frac{D}{\sigma_x}\right) = \exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right) * \frac{x-D}{\sigma_x}$ from the following

$$\begin{aligned} f'\left(\frac{D}{\sigma_x}\right) &= \frac{d\left[\exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right)\right]}{d\left(\frac{D}{\sigma_x}\right)} = \frac{d\left[\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma_x} - \frac{D}{\sigma_x}\right)^2\right)\right]}{d\left(\frac{D}{\sigma_x}\right)} \\ &= \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma_x} - \frac{D}{\sigma_x}\right)^2\right) * \frac{d\left[-\frac{1}{2}\left(\frac{x}{\sigma_x} - \frac{D}{\sigma_x}\right)^2\right]}{d\left(\frac{D}{\sigma_x}\right)} \\ &= \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma_x} - \frac{D}{\sigma_x}\right)^2\right) * \left(-\frac{1}{2}\right) * 2\left(\frac{x}{\sigma_x} - \frac{D}{\sigma_x}\right) * (-1) \\ &= \exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right) * \frac{x-D}{\sigma_x} \end{aligned}$$

Then let $\frac{D}{\sigma_x} = a$ or $D = a\sigma_x$, and take only the first two terms as an approximation"

$$\begin{aligned} f\left(\frac{D}{\sigma_x}\right) &= f(a) + f'(a)\left(\frac{D}{\sigma_x} - a\right) \\ &= \exp\left(-\frac{(x-a\sigma_x)^2}{2\sigma_x^2}\right) + \exp\left(-\frac{(x-a\sigma_x)^2}{2\sigma_x^2}\right) * \frac{x-a\sigma_x}{\sigma_x} * \left(\frac{D}{\sigma_x} - a\right) \end{aligned}$$

$$\text{Let } a=0, f\left(\frac{D}{\sigma_x}\right) = \exp\left(-\frac{x^2}{2\sigma_x^2}\right) + \exp\left(-\frac{x^2}{2\sigma_x^2}\right) * \frac{xD}{\sigma_x^2} = \exp\left(-\frac{x^2}{2\sigma_x^2}\right) * \left(1 + \frac{xD}{\sigma_x^2}\right)$$

Substitute this into $g_r^{pha}(x - D, y)$:

$$\begin{aligned}
g_r^{pha}(x-D, y) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right) \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \cos(\omega(x-D) + \varphi + \Delta\varphi) \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega(x-D) + \varphi + \Delta\varphi) \left(1 + \frac{x D}{\sigma_x^2}\right)
\end{aligned}$$

Then $r_{sr}^{pha} = \iint_{-\infty}^{\infty} g_r^{pha}(x-D, y) I(x, y) dx dy =$

$$\begin{aligned}
&\iint_{-\infty}^{\infty} \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega(x-D) + \varphi + \Delta\varphi) \left(1 + \frac{x D}{\sigma_x^2}\right) I(x, y) dx dy \\
&= \iint_{-\infty}^{\infty} \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega x + \varphi + \Delta\varphi - \omega D) I(x, y) dx dy \\
&+ \iint_{-\infty}^{\infty} \frac{x D}{\sigma_x^2} * \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega x + \varphi + \Delta\varphi - \omega D) I(x, y) dx dy \\
&= A \cos(\alpha + \varphi + \Delta\varphi - \omega D) + \frac{D}{\sigma_x} B \cos(\beta + \varphi + \Delta\varphi - \omega D)
\end{aligned}$$

where $B = \sqrt{B_1^2 + B_2^2}$ and $\beta = \arctan\left(\frac{B_2}{B_1}\right)$

$$B_1 = \iint_{-\infty}^{\infty} \frac{x}{\sigma_x} g_{\cos}(x, y) I(x, y) dx dy \quad \text{and} \quad B_2 = \iint_{-\infty}^{\infty} \frac{x}{\sigma_x} g_{\sin}(x, y) I(x, y) dx dy$$

The simple cell response is:

$$\begin{aligned}
r_s^{pha} &= r_{sl} + r_{sr}^{pha} \\
&= A \cos(\alpha + \varphi) + A \cos(\alpha + \varphi + \Delta\varphi - \omega D) + \frac{D}{\sigma_x} B \cos(\beta + \varphi + \Delta\varphi - \omega D) \\
&= 2A \cos\left(\alpha + \varphi + \frac{\Delta\varphi - \omega D}{2}\right) \cos\left(\frac{\Delta\varphi - \omega D}{2}\right) + \frac{D}{\sigma_x} B \cos(\beta + \varphi + \Delta\varphi - \omega D)
\end{aligned}$$

The quadrature counterpart has a phase difference $\frac{\pi}{2}$ but the same $\Delta\varphi$, so the

quadrature simple cell response is:

$$\begin{aligned}
r_{sq}^{pha} &= r_{sql} + r_{sqr}^{pha} \\
&= 2A \cos\left(\alpha + \varphi - \frac{\pi}{2} + \frac{\Delta\varphi - \omega D}{2}\right) \cos\left(\frac{\Delta\varphi - \omega D}{2}\right) + \frac{D}{\sigma_x} B \cos\left(\beta + \varphi - \frac{\pi}{2} + \Delta\varphi - \omega D\right) \\
&= 2A \sin\left(\alpha + \varphi + \frac{\Delta\varphi - \omega D}{2}\right) \cos\left(\frac{\Delta\varphi - \omega D}{2}\right) + \frac{D}{\sigma_x} B \sin(\beta + \varphi + \Delta\varphi - \omega D)
\end{aligned}$$

The complex cell response is:

$$r = (r_s^{pha})^2 + (r_{sq}^{pha})^2$$

$$\begin{aligned}
&= \left[2A \cos \left(\alpha + \varphi + \frac{\Delta\varphi - \omega D}{2} \right) \cos \left(\frac{\Delta\varphi - \omega D}{2} \right) + \frac{D}{\sigma_x} B \cos(\beta + \varphi + \Delta\varphi - \omega D) \right]^2 \\
&+ \left[2A \sin \left(\alpha + \varphi + \frac{\Delta\varphi - \omega D}{2} \right) \cos \left(\frac{\Delta\varphi - \omega D}{2} \right) + \frac{D}{\sigma_x} B \sin(\beta + \varphi + \Delta\varphi - \omega D) \right]^2 \\
&= 4A^2 \cos^2 \left(\frac{\Delta\varphi - \omega D}{2} \right) + 2A \frac{D}{\sigma_x} B \cos \left(\frac{\Delta\varphi - \omega D}{2} \right) \cos \left(\alpha - \beta - \frac{\Delta\varphi - \omega D}{2} \right) + \frac{D^2}{\sigma_x^2} B^2
\end{aligned}$$

Ignoring the second term (first order in $\frac{D}{\sigma_x}$) and the third item (second order in $\frac{D}{\sigma_x}$)

because the first term (zeroth order in $\frac{D}{\sigma_x}$) is usually larger and keeping only the first

term, an approximation of the complex cell response is expressed as:

$$r = 4A^2 \cos^2 \left(\frac{\Delta\varphi - \omega D}{2} \right)$$

Then the maximum response appears when $\Delta\varphi - \omega D = 0$. Hence the preferred disparity is:

$$D_{\text{pref}}^{\text{pha}} \approx \frac{\Delta\varphi}{\omega}$$

Appendix B. List of Publications

This thesis has contributed to the following publications.

Journal Papers

- [1] **Xuejie Zhang** and Leng Phuan Tay, "A spatial variant approach for vergence control in complex scenes," *Image and Vision Computing*, in press.
- [2] **Xuejie Zhang** and Leng Phuan Tay, "Binocular vergence control using disparity energy neurons," *Journal of Experimental & Theoretical Artificial Intelligence*, in press.
- [3] **Xuejie Zhang** and Alex Leng Phuan Tay, "A physical system for binocular vision through saccade generation and vergence control," *Cybernetics and Systems*, vol 40, no 6, 2009, pp 549-568

Conference Papers

- [1] **Xuejie Zhang** and Alex Leng Phuan Tay, "A phase-shift disparity energy model using image pyramid for vergence control," in *International Conference on Cognitive and Neural Systems (ICCNS09)*, Boston, MA, USA, 27-30 May 2009.
- [2] **Xuejie Zhang**, Alex Leng Phuan Tay, and Alexander Stanza Tan, "Neural classification of objects based on Gabor signature," in *International Joint Conference on Neural Networks (IJCNN08)*. Hong Kong, Aug 2008, pp 893-900.
- [3] Raymond Chun Kiat Ho, Alex Leng Phuan Tay, and **Xuejie Zhang**, "Progressive learning paradigms using the parallel K-iterations fast learning artificial neural network," in *International Joint Conference on Neural Networks (IJCNN07)*. Orlando, USA, 12-17 Aug 2007, pp 552-557.
- [4] **Xuejie Zhang** and Alex Leng Phuan Tay, "Fast learning artificial neural network (FLANN) based color image segmentation in R-G-B-S-V cluster space," in *International Joint Conference on Neural Networks (IJCNN07)*. Orlando, USA, 12-17 Aug 2007, pp 563-568.
- [5] **Xuejie Zhang**, Alex Leng Phuan Tay, and Peiyuan Liang, "Bio-inspired vergence control model for distance estimation of foveal and peripheral objects," in

International Conference on Cognitive and Neural Systems (ICCNS07). Boston, MA, USA, 16-19 May 2007.

- [6] **Xuejie Zhang** and Alex Leng Phuan Tay, "Baseline independent binocular vergence control of 2 DOF Pan-Tilt cameras using a visual cortical model," in *International Conference in Control, Automation, Robotics and Vision (ICARCV06)*. Singapore, 2006.
- [7] **Xuejie Zhang**, Alex Leng Phuan Tay, and Abhinav Saxena, "Vergence control of 2 DOF Pan-Tilt binocular cameras using a log-polar representation of the visual cortex," in *International Joint Conference on Neural Networks (IJCNN06)*. Vancouver, Canada, 2006, pp. 4277-4283.

Under Review

- [1] **Xuejie Zhang**, Alex Leng Phuan Tay, and Abhinav Saxena, "A binocular vision system with attentive saccade and spatial variant vergence Control," *Cybernetics and Systems*, submitted.