Novel Applications of Cooperative and Self-Organizing Neural Networks to Stereo-Disparity Estimation

A Thesis
submitted for the Degree of
Master of Science (Engineering)
in the Faculty of Engineering

A. Jaya Kumar

Department of Electrical Engineering
Indian Institute of Science
Bangalore – 560 012

AUGUST 2005

Acknowledgements

This thesis is by far one of my significant scientific accomplishments till date. And I would like to express my sincere gratitude to the people who made it possible.

I would like to thank my research advisors, Prof. Y. V. Venkatesh and Prof. K. R. Ramakrishnan from the Computer Vision Group, for helping me explore this vast field. I take this oppurtunity to thank the group members: Kumar, Ramya, Sriram, Venkatesh, Sagar, Vijay, Raymond for a memorable and an enjoyable "stay" in the lab. My special thanks to Kumar, Sriram and Sri Krishna Bhat for questioning my thoughts.

I am also grateful to Malik for helping me prepare myself for a career in Physics. As always it is impossible to mention everybody who had an impact on my work.

I believe that in my work I may not have solved the problem, but have made clear what is the problem to be solved. My final conclusion of the work is:

"It is still an open problem!"

Abstract

The thesis deals with the following problem in computational stereo for extraction of three-dimensional scene structure from images:

Given a binocular (or two-camera-stereo) image pair of a physical scene, estimate the depth or distance of all the points in the scene, *as measured from*, say, the line joining the centers of the two cameras?

The basis for attempting to find a solution to the above problem is the fact that a single three-dimensional physical point projects to a unique pair of image points in the two cameras. Therefore, given the two camera images (which are slightly different from each other because of the separation of the two cameras), if we can locate the image pixels that correspond to the same physical point, then it is possible to determine its three-dimensional position. The distance between the two pixels corresponding to the same physical point is called **disparity**. The amount of disparity is inversely proportional to the depth of that point in the scene. The two pixels are said to "correspond" with each other, and are called "correspondence pixels or points".

Stereo-image analysis can be considered as the problem of determining the correspondence points — hence the name, "correspondence problem" (or CP) — from which (i) the disparity at each pixel is obtained directly, and (ii) an image of disparity values (called the disparity map) can be created. Solution to the CP has turned out to be extremely hard as a consequence of, for instance, (a) ambiguities in matching; (b) variations in (i) image intensities of the matching points, (ii) difference in image sizes; (c) blurring caused by defocus; (d) occlusion; (e) specular reflections; (f) lack of textural information; and (g) noise.

The present work is motivated by the remarkable ability of human visual system (HVS) (and of mammals) to perceive depth *effortlessly* from the images formed on their two retinas. These retinal images can be treated as equivalent to the images acquired by two cameras that (i) are separated by a small distance and (ii) have the lenses positioned in such a way as to match the lens-geometry of the human vision system. Human depth perception is based mainly on the processing of the retinal images by the neurons in the brain, using an architecture which is not yet completely understood. Any model which even crudely imitates the performance of the HVS is believed to provide (i) a non-classical framework to deal with stereo-image analysis and, at the same time, (ii) a possible insight into the HVS itself.

A typical example of a natural stereo-pair is shown in Fig. 1: (a) black-and-white images acquired by the left- and the right-cameras; and (b) the anaglyph created from them for viewing by the **red-green goggles**[1] (provided in the thesis). A synthetic pair is shown in Fig. 2. **Automatic depth extraction from such images using artificial neural networks (ANNs) is the goal of the thesis.**

Figure 1: *Tsukuba stereo-pair: (from left to right) left image, right image, Red-Green anaglyph, true disparity map (coded in gray).*

Starting from the pioneering work of Marr and his associates [1, 2, 3], many assumptions have been made, and constraints imposed on the variables, in order to arrive at a solution to CP. Explicit assumptions are: (1) stereo-pairs are epipolar and the epipolar lines are horizontally aligned, i.e., the correspondence points in the two images lie along the same scan lines; (2) the objects have *continuity* in depth and hence in disparity; there is a one-to-one mapping of an image element from one image to the other (*uniqueness*); and (3) there is an ordering of the matchable points. And implicit assumptions are: (1) every pixel or point in one image has a corresponding point in the other; and (2) there is a fixed range for disparity. Typical constraints (relaxed in recent papers) have been: (a) corresponding points have similar intensities / features; (b) surfaces are smooth; and disparity gradient is 1.

If the disparity map can somehow be estimated, physical depth from a disparity map can be computed from a knowledge of camera geometry (focal length of the lenses, horizontal distance between the centers of the lenses, and the distance of the plane of the lenses to the background). *This aspect is not considered in the thesis.*

In view of the vast literature on the problem, no review can be made of all the papers; therefore, in the thesis, we refer to *only* a sampling of relevant and important work. Classical solutions to the correspondence problem have been explored using (i) **area-, feature-**, and **phase-based**; and (ii) **Bayesian approaches**. Amongst the area-based approaches, since a pixel-by-pixel search leads to *ambiguous* matches, a neighborhood

[1]In the absence of a stereoscope, which requires the pair of images as input, the anaglyph provides the same depth perception when viewed through the red-green goggles.

is employed in order to improve the match reliability. In the cooperative algorithms of [4, 1], the initial estimate of match obtained by block-comparison is improved upon, using the continuity and uniqueness constraints. In any *area-based* match strategy, a block of pixels from the left image is compared with different blocks from right image to establish correspondence. Match measures, like sum of squared difference or correlation, are used, and the shifts for which the measure is minimum or maximum, respectively, are declared the disparity values.

In order to increase the reliability of matches, different models for the neighborhood supports have been proposed. For instance, the match support from the local neighborhood gives rise to a **cooperative stereo** algorithm [1] applied to images of binary random-dot textured objects. Based on the variations in the constraints, different models for the neighborhood supports have been proposed. See [5] in which each disparity receives decreasing support from its neighboring elements, depending on the disparity gradient. The algorithm in [5] can handle certain classes of surfaces. For similar results, see [6] where the Pollard-Mayhew-Frisby (PMF) algorithm utilizes disparity gradients (whose magnitudes are constrained to be less than one) in a cooperative manner to solve the stereo disambiguation problem.

In [4], Zitnick *et al.,* extend the cooperative stereo model similar to the one proposed by Marr *et al.,* [1] for gray-scale images. This approach, while retaining all the problems in simple area-based methods, performs better in the presence of noise, and is able to propagate disparity over small *textureless objects* in a textured background.

While area-based processing can provide a dense disparity map, feature-based processing is expected to give an accurate location of discontinuities. Early results in this direction are due to Marr and Poggio [2] who propose a course-to-fine matching strategy using zero-crossings, and in which search regions for a match are refined iteratively. *But* the disparity map obtained is sparse, and requires an interpolation stage [7] to fill in the values of disparity at intermediate points. Further, the detection of feature points is susceptible to noise, and, more importantly, the approach fails to give acceptable results for surfaces with intensity gradient. It is also found that the performance of such algorithms is unsatisfactory at salient features.

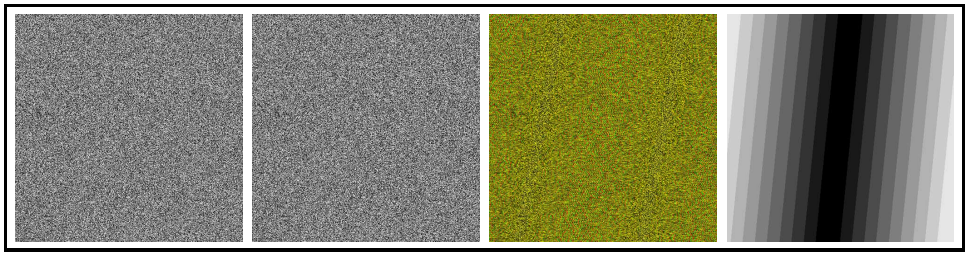
The **phase-based approach** is implicit: disparity is expressed in terms of phase differences in the outputs of local bandpass filters applied to the stereo-pair. However, there exist mathematical limitations with respect to the permissible disparity range.

To summarize, the early algorithms use epipolar constraints, and fail to give a satifactory disparity map in cases where (i) there is vertical disparity in the stereo-pair; and (ii) there are occlusions, as it happens in practice. See [8, 9] and the references therein for recent literature dealing with occlusions. The former reference is based on the Bayesian

framework, and the latter proposes matching outputs of linear spatial filters tuned to a range of orientations and scales. It is found that most of these algorithms do not produce satisfactory disparity maps when applied to standard images [10].

In contrast with the literature, the thesis presents some novel applications of (i) cooperative computation and (ii) self-organizing maps to estimating the disparity map from stereo-image pairs of (i) random dot stereograms, (ii) wire-frame models and (ii) natural images. (*We do not consider multi-camera images nor video sequences of stereo-pairs which can be used to create an enhanced feeling of 3D perception.*)

In the case of the first two (exemplified by Figs. 2 and 3), and partially in the case of the last, (exemplified by Figs. 1 and 4), disparity maps are known which can be exploited to not only validate the proposed algorithms but also compare the results with those of the literature. However, in general, for natural image-pairs (Fig. 5), no such maps are available; therefore, other measures of comparison of the results are required.



(a) (b) (c) (d)

Figure 2: *(a) and (b) Random-dot stereo pair (c) corresponding anaglyph and (d) its true disparity map ($D_{max} = 5$).*

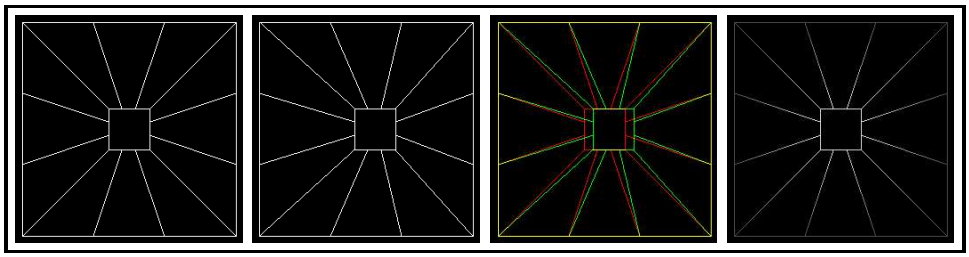


Figure 3: *Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map.*

The **main contributions** of the thesis are:

1. Combination of Cooperative Computation and Supervised Learning

Figure 4: *Left image of the natural stereo-pairs (with known disparity maps): venus, cones, teddy.*

Figure 5: *Pentagon stereo-pairs, no disparity map known: Left image, right image, anaglyph.*

Based on the concept of cooperative stereo, we propose a new model for stereo disparity estimation, using a single-layer artificial neural network (ANN). The main result is that we determine the network weights using synthetic stereo-image data with known disparity maps. The novelty of the approach lies in the fact that the weights characterize the amount of support from each neighbor of the pixel under consideration. The weights so obtained suggest a new model of neuronal excitation-inhibition (for stereo-image analysis) that is distinct from other models.

When the network is applied to random-dot and natural stereo image-pairs, the performance of the proposed network compares favorably with that of recent algorithms. However, its superiority consists in (i) imposing no constraint on disparity gradient; and (ii) a non-iterative computation that is fast and that can be implemented in parallel since the compatibility matrix for each pixel can be constructed separately.

However, in the present form, like most of the other algorithms of the literature. it does not handle occlusion explicitly.

2. Modified Self-Organizing Maps

   A modified self-organizing neural network (SONN) is initialized to, *say*, the right image, and then the amount of deformation required to transform it into the left image is computed.

Its unique characteristic is that the deformation *itself* is a measure of disparity. Some of its special features are: (i) it makes no assumptions with respect to epipolar geometry of the images; and it can take into account salient features (like discontinuous depth values at boundaries and surface wrinkles) and half-occlusions. And when applied to both synthetic and natural stereo-pairs (including wire-frames), performs better than existing algorithms, like the cooperative stereo algorithm of Zitnick and Kanade [4]. Many examples are given to illustrate its superiority and limitations.

Extension of the approach to motion estimation in an image sequence has also been successfully explored.

The cooperative model approach compares favourably with the existing algorithms, but the SONN-based approach is found to be more robust with respect to noise, blurring and quantization than the algorithms of the literature. Further, it is believed that the results obtained from the SONN constitute the *first demonstration* that the disparity map can be easily and efficiently computed with the help of the self-organization principle. The advantages and disadvantages of the methods are also discussed along with some possible guidelines to future research. Typical disparity maps obtained from the application of the two approaches to some of the stereo-image pairs mentioned above are shown in Fig. 6.

The thesis is organized as follows:

- **Chapter 1**: Formulation of the Problem of Computational Stereo.

  Existing approaches to the computation of the disparity map.

  Difficulties related to stereo-image analysis.
- **Chapter 2**: Cooperative Computation

  Proposed approach and the neural architecture.

  Training and testing the network.

  Comparison with the results of the literature.
- **Chapter 3**: Self-Organizing Maps

  Self-organizing neural network and its modification for the CP.

  Experiments with various types of stereo-pairs.

  Comparison with the results of the literature.

  Merits of the proposed approach.

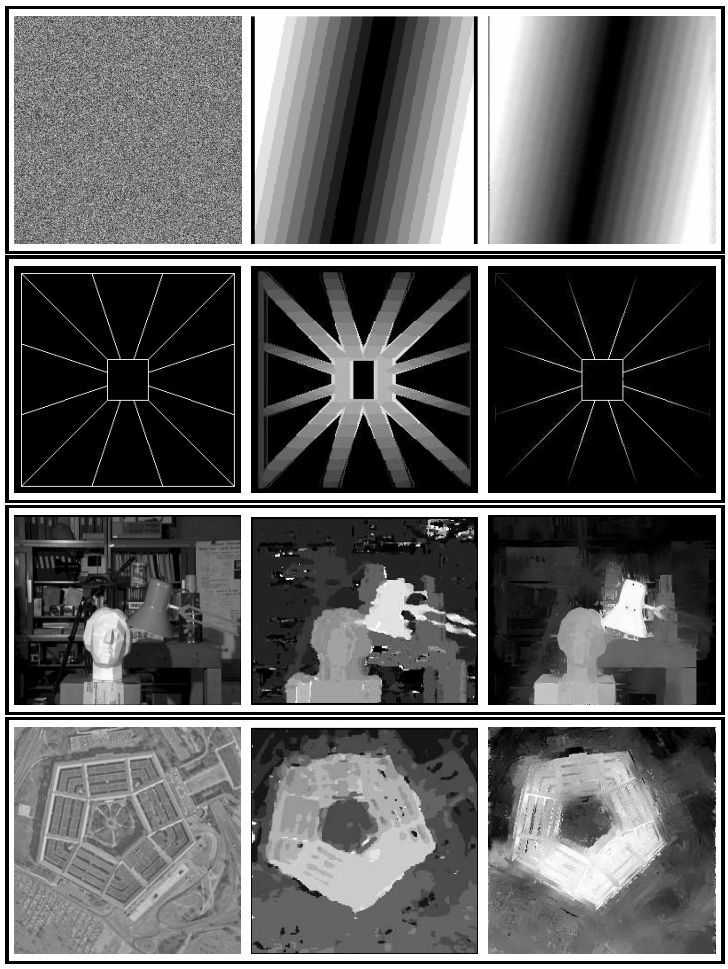


Figure 6: *Estimated disparity maps: (Left to right) (i) Left image of the stereo pair; estimated disparity map using (ii) cooperative; and (iii) self-organizing neural networks.*

- **Chapter 4**: Conclusions.

  Main contributions of the thesis.

  Suggestions for future work: multi-camera stereo, stereo and motion, and real-time stereo-implementation.

Contents

Acknowledgements **i**

Abstract **ii**

1 Introduction **1**
1.1 Stereopsis 1
1.2 Stereo-Pair Acquisition 4
1.2.1 Stereo Geometry 5
1.2.2 Correspondence Problem 6
1.3 Disparity Estimation 7
1.4 Review of Literature 9
1.4.1 Area-based Approaches 10
1.4.2 Feature-based Approaches 14
1.4.3 Cooperative Stereo Models 15
1.4.4 Other Approaches 17
1.5 Generation of Synthetic Stereo-Pairs 18
1.6 Main Contributions 21
1.7 Summary 23

2 Cooperative Model with Supervised Learning **24**
2.1 Proposed Approach 24
2.1.1 Neural Architecture 27
2.1.2 Network Training and Testing 29
2.2 Experimental Results 31
2.3 Summary 40

3 Modified Self-Organizing Neural Network Stereo-Disparity Estimation **41**
3.1 Proposed SONN-based Approach 41
3.2 Analysis of MSOFM 44
3.2.1 Occlusion analysis 47
3.3 Implementation and Results 49
3.3.1 Performance measures for comparison 51
3.3.2 Elementary Stereo-pairs 51
3.3.3 Standard Stereo-pairs 52
3.3.4 Other Stereo-Pairs 59
3.3.5 Merits of MSOFM 62
3.4 Summary 62

4 Conclusions **63**

References **67**

List of Figures

1 Tsukuba stereo-pair: (from left to right) left image, right image, Red-Green anaglyph, true disparity map (coded in gray). iii
2 (a) and (b) Random-dot stereo pair (c) corresponding anaglyph and (d) its true disparity map ($D_{max} = 5$). v
3 Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map. v
4 Left image of the natural stereo-pairs (with known disparity maps): venus, cones, teddy. vi
5 Pentagon stereo-pairs, no disparity map known: Left image, right image, anaglyph. vi
6 Estimated disparity maps: (Left to right) (i) Left image of the stereo pair; estimated disparity map using (ii) cooperative; and (iii) self-organizing neural networks. viii

1.1 Schematic of a stereocope designed by Sir C. Wheatstone(1838) and a modern stereoscope (lab setup). 2
1.2 Left and right image of a stereo pair from the University of Tsukuba. . . . 2
1.3 Construction of an anaglyph from the Tsukuba stereo pair. 3
1.4 Synthetic random dot stereo pair of floating planes and its corresponding anaglyph. 4
1.5 (a) - (b) Stereo pair of a necker cube and (c) its anaglyph. 4
1.6 Stereo pair of a Lambertian sphere and its anaglyph. 5
1.7 Tsukuba stereo pair and the corresponding true disparity map. 5
1.8 Schematic to demonstrate the relation between depth and disparity, and an image of typical lab stereo-setup. 6
1.9 Epipolar geometry of stereo matching. 7
1.10 (Left) Top view of a standard stereo camera setup, (Right) Epipolar geometry for the standard setup 8
1.11 Tsukuba stereo-pair: (from left to right) left image, right image, Red-Green anaglyph, true disparity map (coded in gray). 9
1.12 Left image of the natural stereo-pairs (with known disparity maps): venus, cones, teddy. 9
1.13 Venus stereo-pair: Left image, right image, anaglyph, true disparity map. . 10
1.14 Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map. 10
1.15 Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map. 11

1.16 Left image of natural stereo-pairs (with unknown disparity maps): pentagon, renault. 11
1.17 Pentagon stereo-pairs: Left image, right image, anaglyph. 12
1.18 (Left to right) True disparity map of tsukuba stereo-pair and the corresponding occlusion map. 12
1.19 (Left to right) Left image, right image, anaglyph of Panum's stereo-pair. . 13
1.20 (Left to right) Left image, right image, anaglyph, true disparity map. . . . 13
1.21 Tsukuba stereo-pair with blurred right image: Left image, right image, anaglyph. 14
1.22 Tsukuba stereo-pair with a smaller right image: Left image, right image, anaglyph. 14
1.23 Tsukuba stereo-pair with added noise in right image: Left image, right image, anaglyph. 15
1.24 A stereo-pair with specular reflection : Left image, right image, anaglyph. . 15
1.25 Tsukuba stereo-pair: (from left to right) left image, right image, true disparity map, Red-Green anaglyph 16
1.26 2D support structure of neighborhood for different cooperative stereo models: (a) Marr and Poggio, (b) Grimson, (c) Prazdny, (d) Pollard, Mayhew and Frisby 16
1.27 (Top-Left) 3D view of a fronto-planar object, (Bottom-Left) corresponding true-disparity map, (Top-Right) 3D view of a slant planar object, (Bottom-Right) corresponding true-disparity map. 19
1.28 (Left to right) Red-Green anaglyph of texture mapped fronto-planar object, slant planar object, and the fronto-planar object with repeated texture. In the case of repeated texture, the triangle has a disparity of 10 pixels and the periodicity of the texture is 15 pixels. The perceived depth is equivalent to a disparity of -5 pixels. 19
1.29 (Left to right) Left image, true disparity-map, 3D view, Red-Green anaglyph of (Top-row) LCS stereo-pair and (Bottom-row) WFS stereo-pair. 20
1.30 (Left to right) Left image, right image, anaglyph, true disparity map for Panum's stereo-pair. 21

2.1 Schematic block diagram of the training and test phase. 25
2.2 Structure of compatibility matrix 26
2.3 Neural network structure 27
2.4 (a) and (b) Random-dot stereo pair (c) corresponding anaglyph and (d) its true disparity map ($D_{max} = 5$). 32
2.5 (a) Average neuron responses for different disparities, Estimated disparity map using (b) the network (c) Z-K algorithm. 33
2.6 2D image plot (negative) of weights of different disparity nodes, and their corresponding binary excitation (white) - inhibition (black) plots for $D = 10$ and $D_{max} = 5$. 33

2.7 3D plot of the weights of the zero disparity neuron: (a) general view (b) side view along the principal diagonal (c) the **new Excitation-Inhibition model** for cooperative stereo. Compare with Fig. 1.26. 34
2.8 (a) left image of gray-scale random dot stereo pair (b) corresponding anaglyph; (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network. 34
2.9 (a) Left image of binary random-dot stereo (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network. 35
2.10 (a) left image of wireframe stereo pair (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network. 35
2.11 (a) left image of Tsukuba stereo pair (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network. 36
2.12 (a), (f) right image with noise and blur respectively, (b), (g) anaglyph; and results obtained from/by (c), (h) SSD, (d), (i) Z-K (e), (j) Network. 37
2.13 Left image, anaglyph, true-disparity, SSD, Z-K, and Network results for: (a)-(f) Venus stereo pair, (g)-(l) Cones stereo pair, respectively. 38
2.14 (a), (b) Left and right images of pentagon stereo pair (c) anaglyph (d) estimated disparity map using SSD (e) Z-K algorithm (f) Network, with $D = 10$ and $W = 3$ 39
2.15 Texture-mapped 3D plot of the estimated disparity map of Fig. 2.14 using the network. 39
2.16 (a) Left image of RDS stereo pair (b) True disparity map (c) anaglyph (d) estimated disparity map using SSD (e) Z-K algorithm (f) Network. 39
2.17 (a) Left and right image of Tsukuba stereo-pair with right image vertically shifted by 5 pixels (c) estimated disparity map using SSD (d) Z-K algorithm (e) Network. 40

3.1 Illustration of topology-conserving isomorphic map using SOFM 42
3.2 Block diagram of the MSOFM approach. 44
3.3 Mapping process. 45
3.4 Mapping process before and after occlusion analysis. 45
3.5 Schematic illustration of the updation process. 46
3.6 (Left to right) Left image of a wire-frame stereo, corresponding true disparity map and the estimated disparity map without using the g term . . . 46
3.7 (Left) Image used for both left and right image of stereo pair, i.e. disparity is zero everywhere, (Right) disparity map obtained using standard SOM updation. 47
3.8 (Left to right) left image, right image of the pentagon stereo pair, intensity weight of the network (right image) after 30 iterations with intensity updation, *and* estimated disparity map. 47
3.9 (Top-left) 3D view of the object under consideration, (Bottom-left) Stereo imaging of the object, (Right) Forward and reverse mapping process during disparity estimation. 49

3.10 Occlusion analysis of a fronto-planar textureless object. (Left to right) (S_R) right image, (F_{RL}^{true}) true disparity map, (F_{RL}) disparity map estimated by SOM without occlusion analysis, (O_L) occlusion map, (F'_{RL}) estimated disparity map after occlusion analysis. 50
3.11 (Left to right) Left image of tsukuba stereo-pair, corresponding true-occlusion map (occluded pixels are in white) and the estimated occlusion map. . . . 50
3.12 (Left to right) Left image and the corresponding estimated disparity map using SSD, cooperative stereo, Graph-cut, SOFM for (1st row) FPS with textured object and background, (2nd row) NFPS with textureless object and textured background, (3rd row) FPS with textureless object and background, (4th row) FPS with textured object (periodic texture) and background. 55
3.13 (Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) FPS with binary texture-mapped object and background, (2nd row) Gaussian blurred pair in row 1. 56
3.14 (Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) LCS, (2nd row) WFS. 56
3.15 Natural stereo-pairs with known disparity maps: (Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) tsukuba, (2nd row) venus, (3rd row) cones, (4th row) teddy stereo-pairs. 57
3.16 Natural stereo-pairs with unknown disparity maps:(Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) pentagon, (2nd row) renault stereo-pairs. 58
3.17 (Left to right) Left image, right image, anaglyph, perceived disparity, estimated disparity map using SOFM. 58
3.18 (Left to right) Left image, right image, anaglyph, perceived disparity, estimated disparity map using SOFM. 58
3.19 (Left to right) Left image, right image, anaglyph, true-disparity map, estimated disparity map using MSOFM (1st row) Vertical gradient, (2nd row) horizontal gradient. 59
3.20 (Left to right) Original right image, altered right image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, MSOFM for (1st row) right image with a vertical shift of 5 pixels, (2nd row) right image 90 % of the original height. 59
3.21 (Left to right) Original right image, right image with noise and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) right image with 10% impulse noise (2nd row) right image with Gaussian ($\sigma = 5$) blur. 60

3.22 (Left to right) Original right image of the Tsukuba of stereo pair, the contrast enhanced version, disparity estimate of SSD, cooperative stereo, Graph-cuts and MSOFM. 60
3.23 A stereo-pair with specular reflection : Left image, right image, anaglyph, estimated disparity. 61
3.24 (Left to right) Frame 1 and frame 2 of a video sequence, horizontal and vertical disparity maps. 61

Chapter 1

Introduction

Outline: The problem of stereo-image analysis is formulated, and, after a brief survey of literature, the motivation for the contents of thesis is presented. Finally, the main contributions of the thesis are highlighted.

1.1 Stereopsis

It is well known that humans (in common with mammals) are gifted with stereo vision *thanks* to their two eyes. More explicitly, the images of scenes in the physical world, formed on two retinas, are *apparently* somehow combined in the human brain — the actual process being not yet understood — leading to what is generally called **stereo-perception** of three-dimensional objects in the scene. This ability of a *normal* human visual system to extract depth from the images of the physical world projected on the two retinas is called **stereopsis**.

Many attempts have been made to model stereo-perception, i.e., creation of a system with two cameras (to take the place of the two human eyes) *interfaced with* a computer to imitate this aspect of human vision. Note that each camera, like each human eye, projects the three-dimensional world onto a two-dimensional surface (plane).

Figure 1.1 shows a schematic of an instrument, called the **stereoscope**, with the help of which images from independent sources can be projected onto each eye. The two images shown in Fig. 1.2 are mounted on the stereoscope. When we see the images through the eye-piece of the stereoscope, we vividly perceive depth of different objects in the image. Now, we close one eye, perception of depth is lost. *The conclusion is that our two eyes help us perceive depth.*

The images in Fig. 1.2 have been captured using two cameras that are displaced along the horizontal direction, similar to the two eyes of a human. The images so obtained are called **stereo-pairs**. Another way of viewing stereo images is with an **anaglyph** which is an image formed by displaying the left-image of the stereo pair in red and the right-image

in green overlaid on it (see Fig. 1.3). When we view the anaglyph[1] through a pair of goggles having a red filter for the right eye and a green filter for the left eye, we achieve stereopsis.

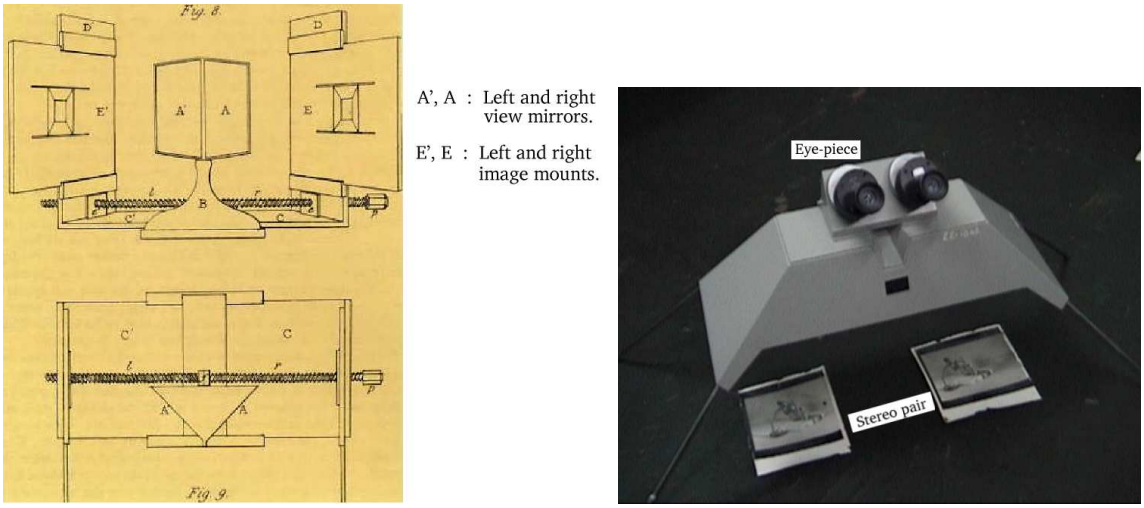


Figure 1.1: *Schematic of a stereocope designed by Sir C. Wheatstone(1838) and a modern stereoscope (lab setup).*

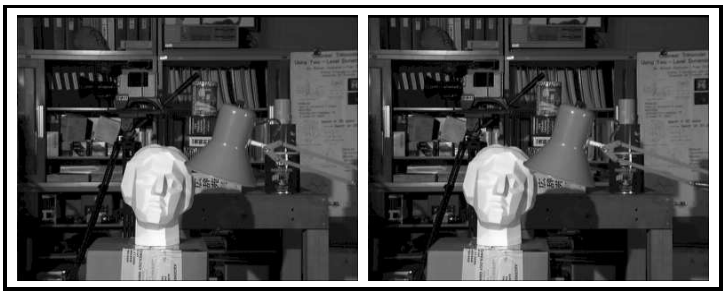


Figure 1.2: *Left and right image of a stereo pair from the University of Tsukuba.*

It is remarkable that stereopsis takes place prior to recognition of any known objects or patterns in the stereo-pair, i.e., no *monocular cues* are needed in the stereo-pair. This was established empirically by Aschenbrunner [11], way back in 1954, with the help of **random stereograms**, which are stereo-pairs of images of objects cluttered with leaves, resulting in no easily discernible object in them when seen separately. But when the stereo-pair is seen through a stereoscope, the objects stand out. In 1971, Julesz [12] synthetically created **random-dot stereograms** by shifting sections of image filled with random dots (see Fig. 1.4). This demonstrated that the human brain is able to perceive

[1]Note that stereoscope requires the pair of images as input. However, the anaglyph provides the same depth perception when viewed through the red-green goggles.

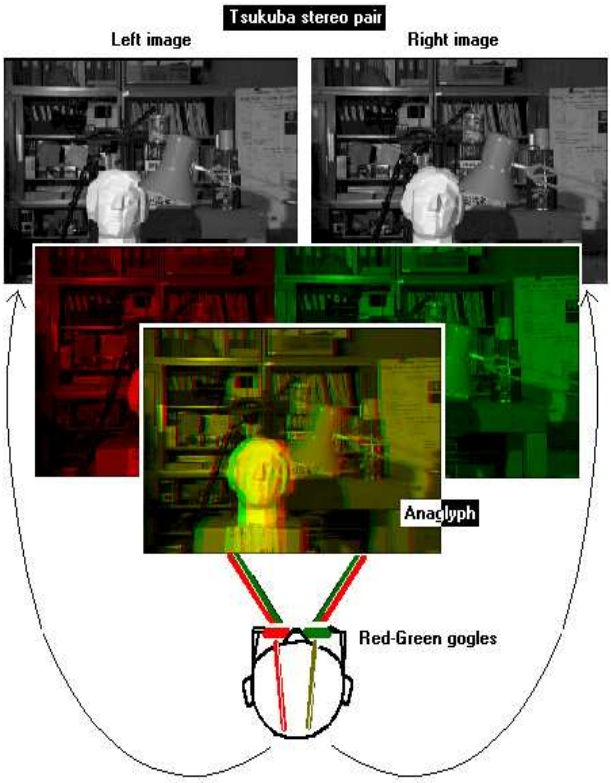


Figure 1.3: *Construction of an anaglyph from the Tsukuba stereo pair.*

depth **independent of monocular cues**. In this context, the shifts in the *corresponding points* (or pixels) in the two images are termed **disparities**.

The thesis deals with the problem of how to compute these disparities from the two images of a scene in order to extract depth information about the objects in the scene. Applications include mobile robotic system with vision and prosthetic devices for the blind.

Remark: It should be noted here that depth perception from stereopsis is quite distinct from that of perspective or shape from shading. In fact, in certain situations where other methods fail to give a unique depth perception, stereopsis can establish uniqueness in depth. For instance, Figs. 1.5 (a) and (b) show the left and right images of a **Necker** cube. When we see them separately, we admit two distinct interpretations of a cube, leading to non-uniqueness in perception[2]; but when we see its anaglyph, Fig. 1.5 (c), through red-green goggles, we perceive *only* one cube. Similarly, Fig. 1.6 shows a Lambertian surface of a sphere, illuminated by a point source. On inverting the image the image appears to be a concave surface. However, in the corresponding anaglyph in Fig. 1.6 the interpretation is invariant to orientation.

[2]This is called *Necker cube reversal.*

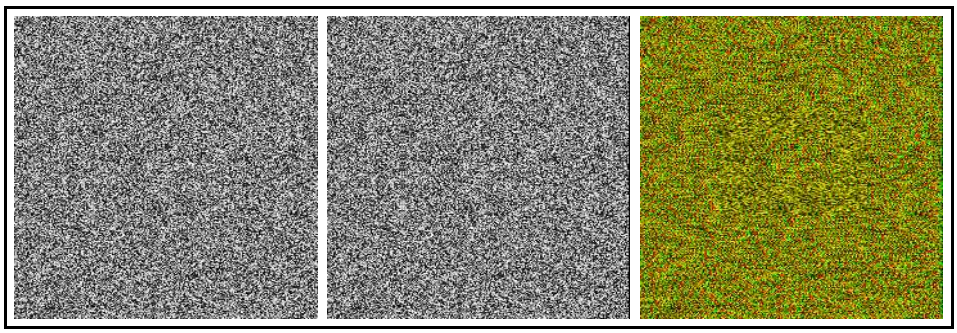


Figure 1.4: *Synthetic random dot stereo pair of floating planes and its corresponding anaglyph.*

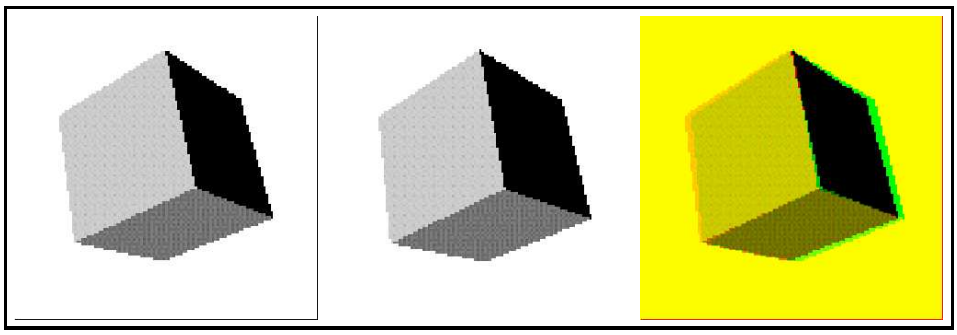


Figure 1.5: *(a) - (b) Stereo pair of a necker cube and (c) its anaglyph.*

1.2 Stereo-Pair Acquisition

Stereo images are *normally* obtained from two identical cameras which are horizontally displaced (*typically*) by 6.5 cm. The cameras are synchronized so that they capture the image simultaneously, thereby avoiding **Pulfrich effect**[3]. Depth in the scene can be estimated from disparity which, *by itself*, gives only the *relative depths* of objects in the image, and the exact relationship between the two depends on the stereo setup. In the next section, we analyse this relationship, assuming standard geometrical constraints that enable us to estimate the actual 3D coordinates of each pixel in the image.

The goal of stereo-image analysis is to estimate the disparity value at every pixel of either the left or the right image. The matrix array giving the disparity values at every pixel is called the **disparity map.** Fig. 1.7 shows a stereo pair and the corresponding disparity map in which each pixel intensity is proportional to the disparity at the corresponding pixel in the left image.

[3]The Pulfrich Effect is an optical illusion in which an object which is moving in a plane seems to move out of that plane.

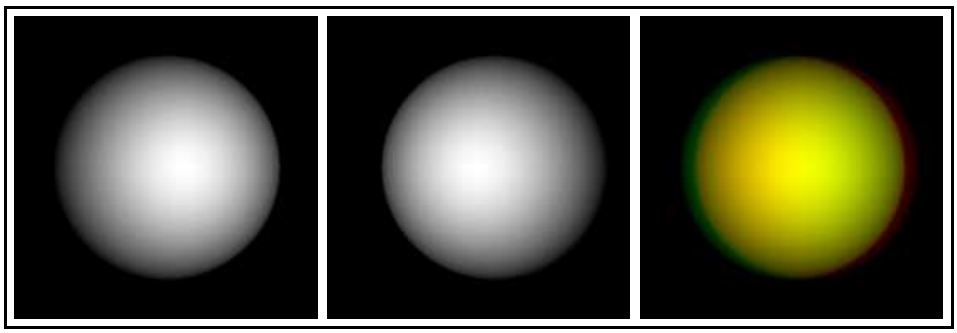


Figure 1.6: *Stereo pair of a Lambertian sphere and its anaglyph.*

Figure 1.7: *Tsukuba stereo pair and the corresponding true disparity map.*

1.2.1 Stereo Geometry

We assume that the two cameras are (a) identical, i.e., their focal lengths and apertures are the same; and they are synchronized while capturing images of scenes; and (b) placed on a rig such that they can be displaced in the horizontal direction by any desired amount. They can also be positioned such that their optic axes are (i) parallel or (ii) verged. In the former case, we encounter only positive disparity whereas in the later the image will have both positive and negative disparity, based on the position of the objects. The cameras may also be at different heights or be rotated along the optic or the horizontal axis. Analysis of stereo-pairs obtained from such setups is not considered in general here. In fact, stereo-pairs captured under general geometric conditions are (mathematically) transformed to the case of parallel camera optic axes and image planes by suitable **calibration**[4] before analysis.

Figure 1.8 shows the top view schematic of a standard stereo setup. The camera centers, C_L and C_R, are separated by a distance, D; I_L, I_R are the image planes of the left and right cameras; and the image planes are at a distance f from the camera centers. The centers of the left and right image planes are L and R. If A and B are points in

[4]This involves estimation of the rotation and translation matrices, and then D, f (see below), so that the images are transformed to those that can be obtained by the standard setup.

the 3D space along the optic axis of the left camera, then A_L, B_L and A_R, B_R are the corresponding projections onto left and right image planes. Since A_L and B_L fall onto L, the disparity of points A and B are RA_R and RB_R. Observe that for $LB > LA$, $RB_R < RA_R$. Using the property of similar triangles, the depth for a given disparity d is found to be given by the algebraic relation,

$$z = \frac{D \times f}{d} - f. \tag{1.1}$$

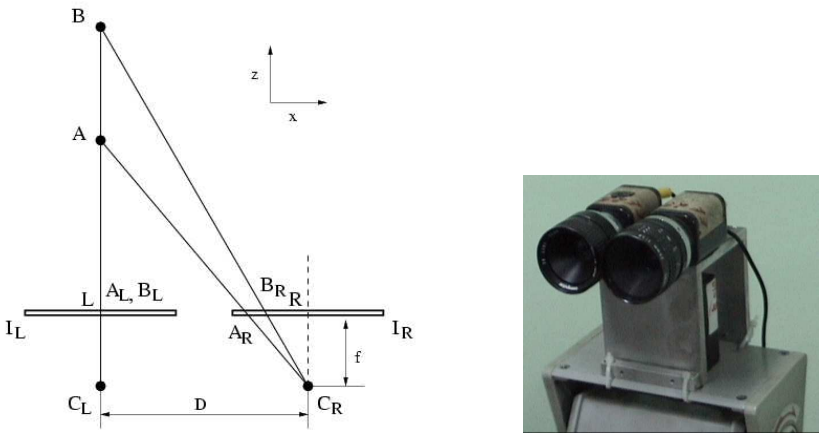


Figure 1.8: *Schematic to demonstrate the relation between depth and disparity, and an image of typical lab stereo-setup.*

For a given camera-setup, calibration involves determining D and f in order to map disparity to physical depth. However, since disparity is fundamental (to the computation of the depth map), we focus *only* on its estimation.

1.2.2 Correspondence Problem

A point in the physical world is mapped onto a *single* point (or pixel) in each of the stereo images. The problem of finding the pixels in the two images that correspond to an arbitrary point in the physical world is fundamental to stereo analysis. It is, in fact, called the **correspondence problem** (CP). Its solution gives the difference (or disparity) in locations of the pixels in the images which, by invoking (1.1), leads to an estimation of the depth of the physical point. Since any solution (to CP) involves searches for matches in a region (i.e., 2D space) and hence is computationally complex, it has been the practice to make a certain assumption on the stereo geometry which is explained below.

See Fig. 1.9 in which all the points along AB project onto point P in the left image plane. In the right image plane, they project onto a straight line segment $A'B'$, called

the **epipolar line.** This implies that the *correspondence point* for P will lie only along $A'B'$. For the standard stereo setup that we considered in the last section — with (i) the cameras at the same height and (ii) the optic axes parallel to each other — the epipolar line lies along the *same scan line.* Hence the search area for correspondence points is reduced to a search along the same scan-lines in the stereo pair. In fact, most of the algorithms in the literature exploit this **epipolar constraint** to reduce computational complexity.

Since, in practice, the camera axes may not be parallel, calibration has to be carried out to estimate the external geometrical parameters which are then used to rectify the images. *This aspect is not considered in the thesis.*

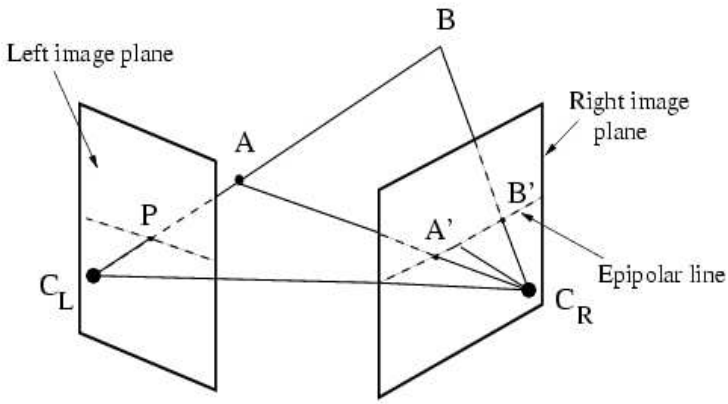


Figure 1.9: *Epipolar geometry of stereo matching.*

1.3 Disparity Estimation

As explained above, computational stereo refers to the problem of estimating the depth of objects in a physical scene from multiple 2D images captured from different viewpoints. In the case of images taken by two cameras with parallel optic-axes and displaced perpendicular to the axes (Fig. 1.10), the depth can be extracted from the difference in the positions of pairs of pixels in the two images that **correspond** to a single physical point.

The stereo-pairs used in our experiments serve as (i) illustrations for the technique proposed in the paper; and/or (ii) benchmarks for comparison with the results of the literature; or/and (iii) counterexamples to provide motivation for further work. More

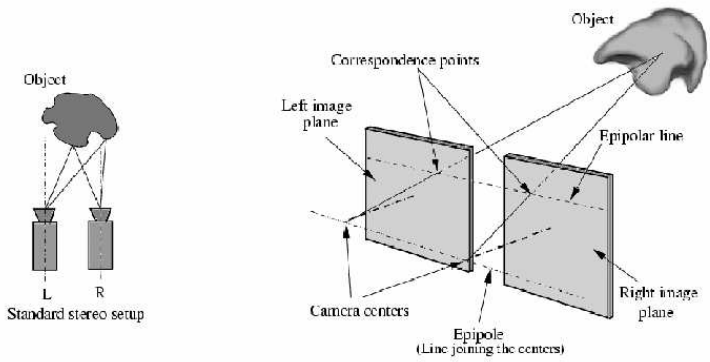


Figure 1.10: *(Left) Top view of a standard stereo camera setup, (Right) Epipolar geometry for the standard setup*

importantly, they serve to formulate a **realistic** problem of stereopsis. In order to facilitate viewing without the help of a stereoscope, we provide anaglyphs of stereo-pairs, (the left image is in red, and the right in green) which can be seen by using red-green goggles (provided in the thesis)

Here are some examples of stereo-pairs: (1) Tsukuba stereo-pair (Fig. 1.11) and the Red-Green anaglyph[5]. (2) Database stereo-pairs (Fig. 1.12) from "http://www.middlebury.edu/stereo/" with true disparity maps (obtained from structured light [13]): "venus", "cones", "teddy". (3) Wire-frame model (with known disparity map): (4) Random-dot stereograms (with known disparity maps): (5) Natural images for which correct disparity map is unknown (Fig. 1.16). (6) Occlusion-affected images[6] (Fig. 1.18). (7) Half-occlusion[7] (with known disparity map) (Fig.1.19). (8) Crease with known disparity map (Fig. 1.20). (9) One image is blurred. (10) One image is slightly smaller than the other. (11) Noisy pair. (12) Specular pair (Fig. 1.24).

The problem of stereo analysis can be formulated as follows: **Given a stereo-pair of images belonging to any one class or a combination of classes described above, determine the disparity of each pixel in the image.**

Such stereo-pairs are, *in general*, associated with the following difficulties during an estimation of disparity at every pixel:

[5]Courtesy University of Tsukuba

[6]Occlusions are regions in space that cannot be seen in both the images.

[7]There are points in one image that have no corresponding matches in the other image. All stereo-pairs with discontinuity in disparity have half-occlusions.

Figure 1.11: *Tsukuba stereo-pair: (from left to right) left image, right image, Red-Green anaglyph, true disparity map (coded in gray).*

Figure 1.12: *Left image of the natural stereo-pairs (with known disparity maps): venus, cones, teddy.*

(1) Noise and illumination changes (and specularity) as a result of which the feature values for the corresponding points in the left and right images differ. (2) Lack of features in large regions. (3) Presence of salient features: discontinuities in depth at object boundaries, discontinuities in surface orientation ("creases"), and steeply sloping surfaces. (4) Occlusion. (5) Half-occlusion.

1.4 Review of Literature

Psycho-visual work on the problem of human stereopsis has been active for a very long period. However, its impact on the development of its computational model, by way of quantification of the human abilities, seems to be visibly insignificant. On the other hand, the literature is vast on a *pure* Marr-paradigm-inspired [3], computational approach to stereopsis that covers a period of about 35 years or so, and cannot be comprehensively reviewed here. See [14, 10] for surveys and for related references. For lack of space, only the most relevant and recent results are briefly examined here.

Starting from the pioneering work of Marr and his associates [1, 2], many assumptions have been made, and constraints imposed on the variables, in order to arrive at a solution to the above problem. Explicit assumptions are: (1) Stereo-pairs are epipolar and the epipolar lines are horizontally aligned, i.e., the correspondence points in the two images lie along the same scan lines (see Fig. 1.10); (2) the objects have *continuity* in depth and

Figure 1.13: *Venus stereo-pair: Left image, right image, anaglyph, true disparity map.*

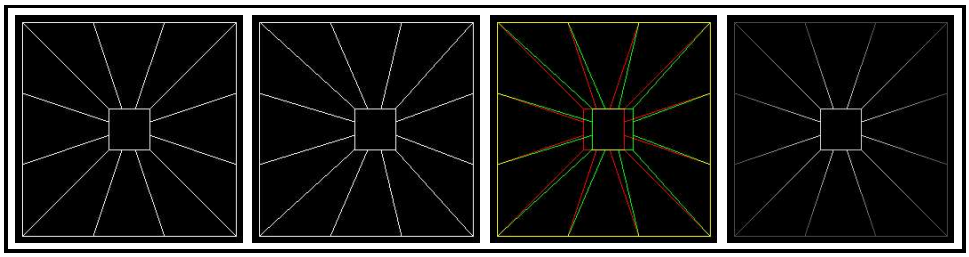


Figure 1.14: *Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map.*

hence in disparity; there is a one-to-one mapping of an image element from one image to the other (*uniqueness*); and (3) there is an ordering of the matchable points ([15]). And implicit assumptions are: (1) every pixel or point in one image has a corresponding point in the other; and (2) there is a fixed range for disparity. Typical constraints (relaxed in recent papers) have been: (a) corresponding points have similar intensities / features; (b) surfaces are smooth; and disparity gradient is 1.

Classical solutions to the correspondence problem have been explored using (i) **area-, feature-**, and **phase-based**; and (ii) **Bayesian approaches**.

1.4.1 Area-based Approaches

In the early area-based algorithms, disparity was assumed to be nearly constant; it was determined by first correlating a window of points in the left image with a window in the right image, and then choosing — for each point in the left image — the disparity which gave rise to the best correlation. Match measures, like sum of squared difference or correlation, are used, and the shifts for which the measure is maximum are declared the disparity values. See [16] for an evaluation of block-based methods for disparity estimation. Area-based algorithms [17] provide a dense disparity field, and are used for disparity estimation in the context of stereo-image coding.

In order to increase the reliability of matches, different models for the neighborhood

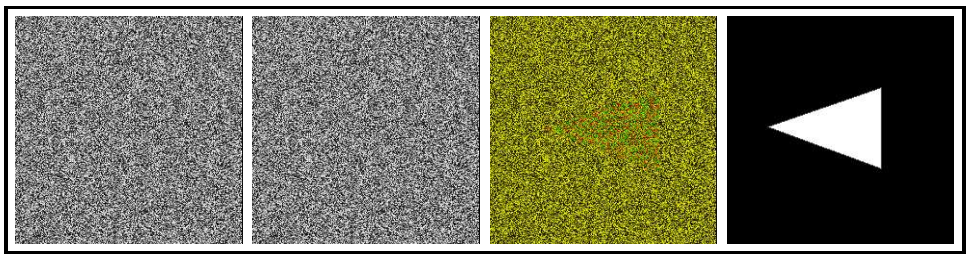


Figure 1.15: *Wire-frame stereo-pair (with known disparity maps): Left image, right image, anaglyph, true disparity map.*

Figure 1.16: *Left image of natural stereo-pairs (with unknown disparity maps): pentagon, renault.*

supports have been proposed. For instance, the match support from the local neighborhood is used to evaluate the correctness of the match. Such a procedure gives rise to a **cooperative stereo** algorithm, originally suggested by Marr and Poggio [1] in combination with continuity and uniqueness constraints for binary random-dot textured objects. In this algorithm, the neighboring pixels are forced to take similar disparity values, and each pixel in one image is to have a unique match in the other image of the stereo-pair. Based on the variations in the constraints, different models for the neighborhood supports have been proposed (see [5]). In the other model of neighborhood support [5], each disparity receives decreasing support from its neighboring elements, depending on the disparity gradient. The algorithm in [5] can handle (i) smooth surfaces that are not fronto-parallel (such as tilted planes), and (ii) transparent objects. For similar results, see [6] where the Pollard-Mayhew-Frisby (PMF) algorithm utilizes disparity gradients (whose magnitudes are constrained to be less than one) in a cooperative manner to solve the stereo disambiguation problem.

In [4], Zitnick *et al.* extend the cooperative stereo model similar to the one proposed by Marr *et al.* [1] for gray-scale images. This approach, while retaining all the problems in simple area-based methods, performs better in the presence of noise, and is able to propagate disparity over small *textureless objects* in a textured background.

Block matching methods, on the other hand, perform well on textured objects but do not give correct disparities for textureless objects. Kanade *et al.* [18] employ an

Figure 1.17: *Pentagon stereo-pairs: Left image, right image, anaglyph.*

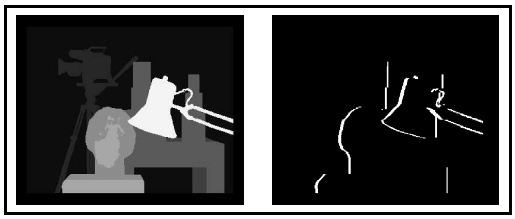


Figure 1.18: *(Left to right) True disparity map of tsukuba stereo-pair and the corresponding occlusion map.*

adaptive window/block to handle the textureless regions. But the block-matching gives rise to artifacts in the case of narrow objects (like wire-frames). When applied to repeated textures, based on the coding technique adopted for deciding the best match, the results show considerable variations. But the major shortcoming of such approaches is that the measures used are very sensitive to noise and to occlusion.

Authors have also used some *ad hoc* measures to improve disparity computation, like local transforms for establishing correspondence [19], and graph-cuts [20, 21]. The latter measure seems to perform well on textureless objects in a textured background. Another approach is to use the stereo constraints to impose some structure on the disparity map.

The relaxation-based approach [22] is similar to the cooperative stereo, except that the energy functions, used for refining the match, are defined explicitly. The match measure used for the initial estimate can be obtained by the area-based methods. For different disparity values, the match measures of each pixel are used to define a probability density function such that the disparity corresponding to the best match has the highest probability. The probability in the next iteration is updated based on continuity and uniqueness conditions. The method does not take into account occlusions, leading to false mapping; and the boundaries in the disparity map do not accurately depict the objects. Moreover, the method fails to propagate disparity into textureless regions, and gives *blocky disparities* for wire-frame images.

In the *dynamic programming* (DP) approach [23], instead of refining the match values, the disparity variation is modeled by a discretized curve such that it minimizes a

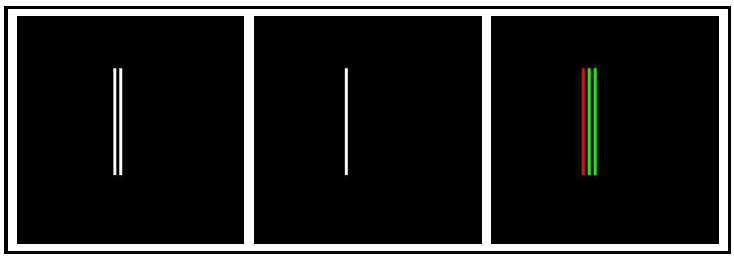


Figure 1.19: *(Left to right) Left image, right image, anaglyph of Panum's stereo-pair.*

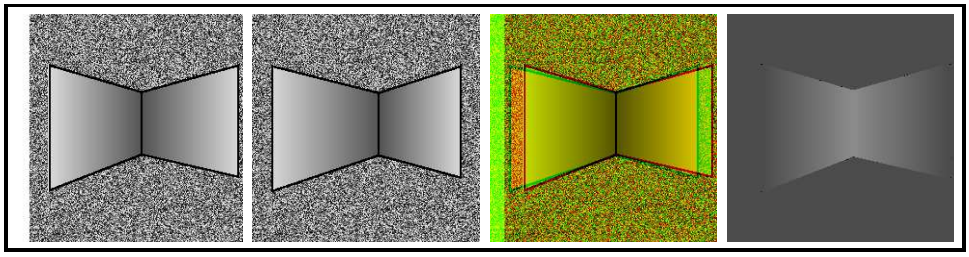


Figure 1.20: *(Left to right) Left image, right image, anaglyph, true disparity map.*

cost function expressed in terms of the match measure and smoothness. However, the method does not employ a uniqueness constraint but makes use of an *ordering constraint*, according to which the corresponding points in both the images are in the same order, and exist only along the scan lines. The ordering helps in propagating disparity in the textureless regions, but fails in the presence of narrow objects. It is highly sensitive to noise and occlusion, and a false match may propagate and destroy other good matches. Moreover, the disparity is always an integer value, and the continuity constraint is applied only along the epipolar line.

In the early DP algorithms of [24, 25], two scan-lines are matched on the left and right images, instead of matching individual pixels or blocks. As a consequence, the method allows global consistency constraints to be imposed. But these constraints are imposed along scan-lines and not across them. The result is poor estimates of across edges.

Alternately, in the *graph-cut*-based approach [26], the stereo problem is cast as one of finding maximum flow graph in a combinatorial optimization framework. Here the ordering constraint is discarded in favor of a local coherence constraint. The basic idea is to fit a 2D surface in the 3D disparity space, such that they obey the constraints. Different graph structures have been proposed [20], [21] which optimize efficiently. According to these authors, the algorithm performs well on textureless objects in a textured background.

As far as the matching operation is concerned, measures like sum of squared difference (**SSD**) or correlation are used, and the shifts for which the measure is minimum or

Figure 1.21: *Tsukuba stereo-pair with blurred right image: Left image, right image, anaglyph.*

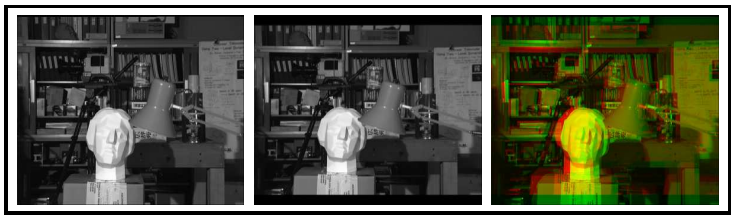


Figure 1.22: *Tsukuba stereo-pair with a smaller right image: Left image, right image, anaglyph.*

maximum lead to the disparity-map. This approach gives good results for dense textured synthetic images like the random dot stereograms (RDS) [27, 10], and can be improved by normalizing the match measures. Match measures for different disparity values can be used to produce a 3D projective transformation with (x, y, d) as the coordinates. The match value at (x, y, d) corresponds to the match measure between the blocks centered at (x, y) and $(x + d, y)$ of the left and right images, respectively. In regions of low texture, such measures lead to ambiguity in matches. They are also susceptible to noise and intensity variations.

1.4.2 Feature-based Approaches

While area-based processing can provide a dense disparity map, feature-based processing is expected to give an accurate location of discontinuities. The robustness of the human visual system to noise and intensity variations has prompted the use of edges and zero crossings to obtain reliable matches. But the disparity map thus obtained is sparse and constrained to the edges/zero-crossings. And, in the case of dense textures, like the RDS, the matching process is ambiguous.

Early results in this direction are due to Marr and Poggio [2] who propose a course-to-fine matching strategy using zero-crossings, and in which search regions for a match are refined iteratively. *But* the disparity map obtained is sparse, and requires an interpolation stage [7] to fill in the values of disparity at intermediate points. Further, the detection of feature points is susceptible to noise, and, more importantly, the approach fails to

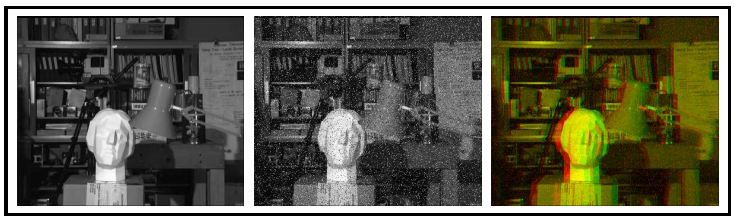


Figure 1.23: *Tsukuba stereo-pair with added noise in right image: Left image, right image, anaglyph.*

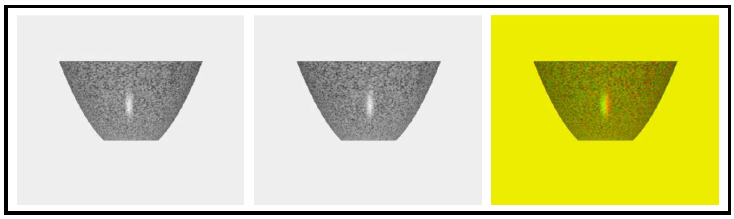


Figure 1.24: *A stereo-pair with specular reflection : Left image, right image, anaglyph.*

give acceptable results for surfaces with intensity gradient. It is also found that the performance of such algorithms is unsatisfactory at salient features.

Assuming the scene consists of smooth surface patches, Lin and Tomasi [28] model the disparities (in those patches) by a continuous-valued spline. They iteratively estimate disparities and extents of these patches in the framework of energy minimization. It is found that such a procedure can reliably estimate disparity in textureless regions and non-frontal planes, but fails in fine textured images of the type of random-dot stereograms and images with smooth intensity gradient (as in Lambertian images). Moreover, the segmentation has to be independent of the contrasts of the images to produce similar segments in both the images.

For textured outdoor scenes and highly regular man-made objects, Cochran and Medioni [29] present a stereo-vision system that integrates area- and feature-based primitives. They employ area-based cross correlation along with an ordering constraint and a weak surface smoothness assumption to produce an initial disparity map. At depth discontinuities, edge information is introduced to reduce the smoothing caused by cross correlation. Such a procedure facilitates detection of depth and orientation discontinuities in the images.

1.4.3 Cooperative Stereo Models

The reliability of *area-* and/or *feature-based* matches is improved by employing the support from the local neighborhood to evaluate the correctness of the match (and hence

Figure 1.25: *Tsukuba stereo-pair: (from left to right) left image, right image, true disparity map, Red-Green anaglyph*

the name). All the models [30] use the **epipolar constraint**. Assuming continuity and uniqueness of matches, Marr and Poggio [1] propose a model shown in Fig. 1.26 in which the support in the local neighborhood is plotted (in 2D) as a function of change in (i) disparity (y-axis) and (ii) distance along the epipolar line (x-axis).

In the darker region, a match at a point reduces the reliability of the match point under consideration. Therefore, such a region has an *inhibitive* effect. The gray region acts as *excitation*, increasing the reliability of the match. The continuity constraint assumes that the depth is constant, and therefore the matches lying along the same disparity planes are excitatory. Assuming each pixel in one image has an unique correspondence pixel in the other image, the matches which use the same pixels are inhibited. Grimson [7] relaxes the continuity constraint, and allows a small change in disparity, so as to take into account small gradients in disparity. The figure also shows the model proposed by (i) Pollard *et al.* [6] who use constraints on the disparity gradient in the local neighborhood and (ii) Prazdny [5] who invokes a coherence principle. See Hakkarainen *et al.* [30] for more details and for a comparative study of these models.

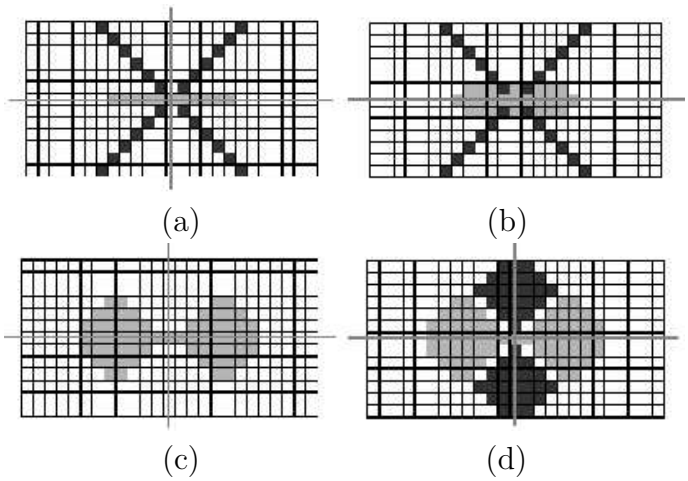


Figure 1.26: *2D support structure of neighborhood for different cooperative stereo models: (a) Marr and Poggio, (b) Grimson, (c) Prazdny, (d) Pollard, Mayhew and Frisby*

The support structures are then used as a mask to convolve in the 3D disparity-space. This helps enhance the match-measure based on the assumptions.

From a computational point of view, for all the models, the amount of excitation/inhibition support contributed by each match in the neighborhood of a pixel is based on heuristics. Khotanzad *et al.* [31] use the match values from the excitation-inhibition regions defined by Marr and Poggio's model as input to train a multilayer-perceptron. They show that the network develops excitation-inhibition weights as predicted by Marr and Poggio's model. See [32] for an extension of a similar approach to gray scale images using the SSD for initial estimate to train a recurrent network. In both [31, 32], it is to be noted, the authors use Marr and Poggio's cooperative stereo model, and consider only the match values lying along the excitation and inhibition regions given by the model.

1.4.4 Other Approaches

In contrast with the above explicit approaches, the **phase-based approach** of [33] is implicit: disparity is expressed in terms of phase differences in the outputs of local bandpass filters applied to stereo-pair. However, there exist mathematical limitations with respect to the permissible disparity range. For an interesting biological counterpart, see [34].

It should be added here that the early algorithms[8] use epipolar constraints, and fail to give an acceptable disparity map whenever there is vertical disparity in the stereo-pair as it happens in practice. In such cases, the camera axes may not be parallel, and calibration has to be carried out to estimate the external geometrical parameters which are then used to rectify the images to be able to employ epipolar search operations (for correspondence). Also these algorithms *either* do not explicitly deal with occlusions and other phenomena *or* give unsatisfactory results at occlusions. In fact, it is only recently that researchers have begun to address occlusion. See [8, 35, 9] and the references therein. A complex graph proposed by Kolmogorov *et al.* [36] handles occlusions by invoking the uniqueness constraint. But the approach fails in the case of wire-frame images. Geiger et al. [35], using the Bayesian approach, show that occlusions can help stereo computation by providing cues for depth discontinuities. They introduce a constraint that relates discontinuities in one image with occlusions in the other image. In the course of this procedure, they assume that the corresponding epipolar lines between the two images are known. Disparity discontinuity in one image always corresponds to an occluded region

[8] with the possible exception of Bayesian approaches that require changes in the formulation of the priors. See below.

in the other image, thus leading to an occlusion constraint or monotonicity constraint which restricts the space of possible disparity values, thereby simplifying computations, they find an optimal solution using dynamic programming.

Jones and Malik [9] incorporate some stereo-cues to deal with occlusions, while matching outputs of linear spatial filters tuned to a range of orientations and scales. An alternative theory, based on the Bayesian framework, dealing with stereo and occlusions has been developed by Belhumeur and Mumford (see [8] and the references therein) in order to take care of salient features and half-occlusion. Belhumeur [8] obtains a *maximum a posteriori* (MAP) estimate for the prior model (of scene geometry) which includes not simply depth, but also orientation, discontinuities in depth at object boundaries, and surface creases. Belhumeur proposes that depth, surface orientation, occluding contours, and creases should be estimated simultaneously. To accomplish this, he develops a class of energy functionals (which includes the "weak string", the "weak rod" and the "weak rubber band") that implicitly assumes a prior model that is constructed from the sums of Brownian motion processes and compound Poisson processes. This model (called the hybridized prior) is extended to consider global structures (to take care of highly non-local interactions in the scene geometry).

Remark: No reference is made here to multi-camera images nor video sequences of stereo-pairs which can be used to create an enhanced feeling of 3D perception. In [37], occlusion in two-image stereo is overcome by considering multi-view stereo images to detect occlusion. We do not deal with such images here. See [38] for a survey of literature on (coding related to) stereo from video sequences.

1.5 Generation of Synthetic Stereo-Pairs

In order to evaluate and compare the performances of different stereo algorithms, we first generate certain classes of elementary stereo-pairs which can be used as basis images to reconstruct other stereo-pairs with *non-specular reflecting surfaces.* They help us identify limitations on the use of constraints in the algorithms (for disparity computation), thereby facilitating a classification (as also a comparison) of algorithms based on performance.

(1) **Planar surfaces:** The images are of planar surfaces that are texture-mapped with different types of textures or may be textureless. From the point of view of stereo constraints, planar surface orientation in 3D can be studied using the following two categories (see Fig. 1.27):

(a) **Fronto-planar surfaces (FPS):** These are visible surfaces of objects that are

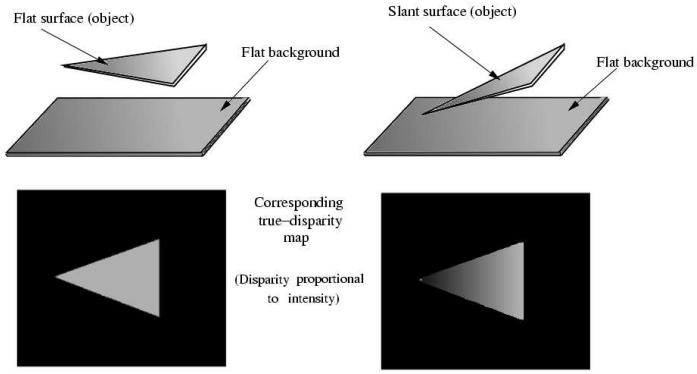


Figure 1.27: *(Top-Left) 3D view of a fronto-planar object, (Bottom-Left) corresponding true-disparity map, (Top-Right) 3D view of a slant planar object, (Bottom-Right) corresponding true-disparity map.*

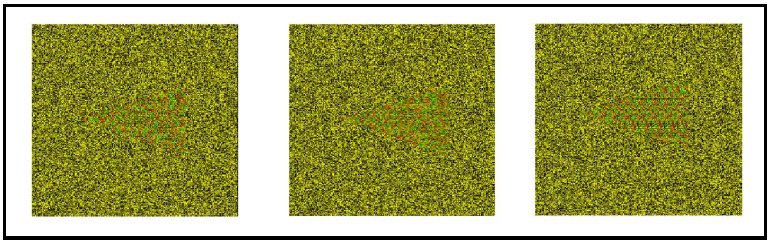


Figure 1.28: *(Left to right) Red-Green anaglyph of texture mapped fronto-planar object, slant planar object, and the fronto-planar object with repeated texture. In the case of repeated texture, the triangle has a disparity of 10 pixels and the periodicity of the texture is 15 pixels. The perceived depth is equivalent to a disparity of -5 pixels.*

parallel to the image plane. Therefore, the whole surface will have a uniform disparity. The corresponding pixels of the object have a one-to-one map. The performance of the algorithms on these stereo-pairs shows the use of continuity constraint. In the case of textureless objects, the ability of an algorithm to propagate disparity can be tested.

(b) **Non-frontal planar surface (NFPS):** These are planar surfaces which are at an angle with the image plane. The surface will have a uniform disparity gradient (Fig. 1.27). Here, the correspondence pixels do not have a one-to-one mapping, hence they violate the uniqueness condition. The results of the algorithms on these stereo-pairs will demonstrate the interaction between continuity and uniqueness constraints.

We use a triangular plane as the object with a fronto-planar background (Fig. 1.27). The results of the algorithms also depend on the types of texture, based on the matching approach used. Since the types of textures are innumerable, we use random-dots with different sparsity to check on matching. In order to evaluate the performance of any

algorithm for planar surfaces, we need to experiment with all combinations of textured and textureless, object-background interfaces.

A special case arises when the texture is periodic. If the periodicity of the texture is n and the actual disparity at a point is d, observations show that the perceived depth is corresponding to the disparity of min $(d \ mod \ n, n - (d \ mod \ n))$ (Fig. 1.28). The fused stereo-pair is also shown in Fig. 1.28.

(2) **Curved surfaces:** Here the objects could be of any shape, but from the stereo point of view the objects considered are convex and hence do not have self-occlusions. Textureless surfaces do not give rise to curved surfaces. On the other hand, Lambertian surfaces (**LCS**) with intensity gradients produce curvatures (Fig. 1.29). We can evaluate (i) the ability of the algorithm to use intensity for matching, and (ii) smoothness criteria. These stereo-pairs do not have extractable features like zero-crossings, and hence the feature-based approaches fail to produce disparity.

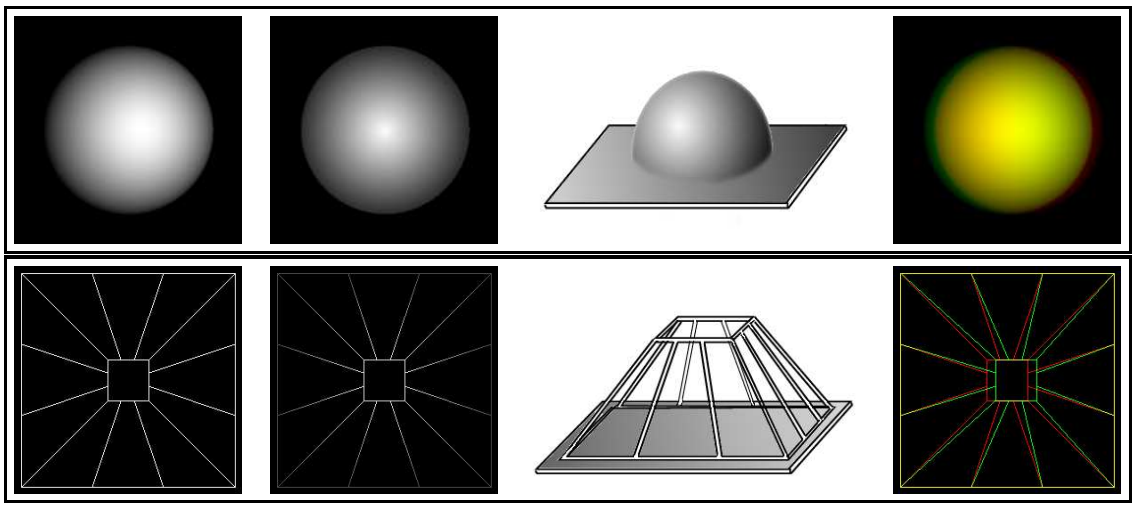


Figure 1.29: *(Left to right) Left image, true disparity-map, 3D view, Red-Green anaglyph of (Top-row) LCS stereo-pair and (Bottom-row) WFS stereo-pair.*

(3) **Objects with occlusion, half-occlusion:** A discontinuity in disparity map along the epipolar line, in the above cases give rise to half-occlusions. To analyze the occlusion handling capability of the algorithm, we use the stereo pair with Panum's limiting case (Fig. 1.30).

(4) **Thin/fine objects (Wire-frames):** One of the most significant features of wire-frame images is that they contain no clues (like shading or texture) which are important pointers to depth information in (natural) monocular images. The wire-frame images (i) have sparse match points, with disparity confined to the lines; and (ii) represent the narrow objects (in the image). Such images help us test the performance of a stereo-algorithm on fine-structured objects.

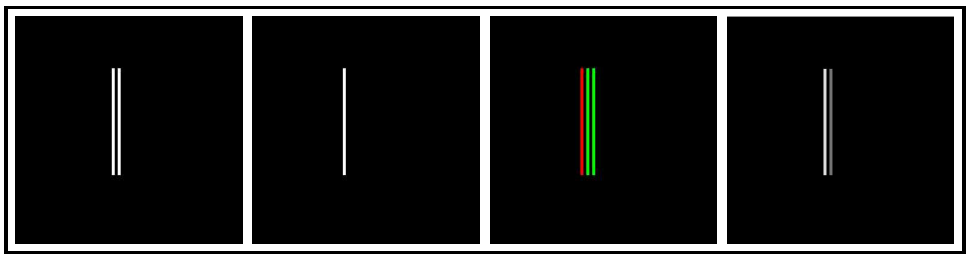


Figure 1.30: *(Left to right) Left image, right image, anaglyph, true disparity map for Panum's stereo-pair.*

1.6 Main Contributions

In contrast with the literature, the thesis presents some *novel* applications of (i) cooperative computation and (ii) self-organizing maps to estimating the disparity map from stereo-image pairs of (i) random dot stereograms, (ii) wire-frame models and (ii) natural images:

1. **Combination of Cooperative Computation and Supervised Learning**

   Based on the concept of cooperative stereo, we propose a new model for stereo disparity estimation, using a single-layer artificial neural network (ANN). The main result is that we determine the network weights using synthetic stereo-image data with known disparity maps. The novelty of the approach lies in the fact that the weights characterize the amount of support from each neighbor of the pixel under consideration. The weights so obtained suggest a new model of neuronal excitation-inhibition (for stereo-image analysis) that is distinct from other models.

   When the network is applied to random-dot and natural stereo image-pairs, its performance compares favorably with that of recent algorithms. Its superiority consists in (i) imposing no constraint on disparity gradient; and (ii) a non-iterative computation that is fast and that can be implemented in parallel since the compatibility matrix for each pixel can be constructed separately.

   However, in the present form, like most of the other algorithms of the literature. it does not handle occlusion explicitly.

2. **Modified Self-Organizing Feature Maps**

   We adopt a strategy that involves *pixel-by-pixel* matches by treating the CP as an estimation of a *mapping* of non-occluded pixels from one image to the other. This idea is implemented by using an isomorphic self-organizing neural network (SONN) with some modifications to include, systematically, stereo constraints (like

occlusions, half-occlusions and salient features). The network is initialized to, say, the right image, and then the amount of *deformation* required to transform it into the left image is determined. The competitive learning approach of the algorithm is *exploited* to find the best match in combination with local cooperation which leads to propagation of disparity. The amount of deformation can be used as the disparity measure.

Its unique characteristic is that the deformation *itself* is a measure of disparity. Some of its special features are: (i) it makes no assumptions with respect to epipolar geometry of the images; and (ii) it can take into account salient features (like discontinuous depth values at boundaries and surface wrinkles) and half-occlusions, (iii) when applied to both synthetic and natural stereo-pairs (including wire-frames), performs better than existing algorithms. Many examples are given to illustrate its superiority and limitations.

Extension of the approach to motion estimation in an image sequence has also been successfully explored.

The cooperative model approach compares favourably with the existing algorithms, but the SONN-based approach is found to be more robust with respect to noise, blurring and quantization than the algorithms of the literature. Further, it is believed that the results obtained from the SONN constitute the *first demonstration* that the disparity map can be easily and efficiently computed with the help of the self-organization principle. The advantages and disadvantages of the methods are also discussed along with some possible guidelines to future research.

With the main results presented in Chapters 2 and 3, the thesis is organized as follows:

- **Chapter 1**: Formulation of the Problem of Computational Stereo.

  Existing approaches to the computation of the disparity map.

  Difficulties related to stereo-image analysis.
- **Chapter 2**: Cooperative Computation

  Proposed approach and the neural architecture.

  Training and testing the network.

  Comparison with the results of the literature.
- **Chapter 3**: Self-Organizing Maps

  Self-organizing neural network and its modification for the CP.

  Experiments with various types of stereo-pairs.

  Comparison with the results of the literature.

  Merits of the proposed approach.
- **Chapter 4**: Conclusions.

  Main contributions of the thesis.

  Suggestions for future work: multi-camera stereo, stereo and motion, and real-time stereo-implementation.

1.7 Summary

After formulating the problem of stereo-image analysis, we have presented (i) a review the relevant literature and (ii) the motivation for the proposed approach based on neural networks. Finally, the main contributions of the thesis are highlighted.

Chapter 2

Cooperative Model with Supervised Learning

Outline: For stereo-disparity estimation, we present a new cooperative model using an artificial neural network (ANN)[1]. We assume that sample stereo-image pairs are available with known depth- (or, equivalently, disparity-) maps which are used as training datasets. A supervised learning strategy determines the weights of the network, specifying the amount of support from each neighbor. The *existing* cooperative stereo algorithms use different support structures, with excitation-inhibition weights based on certain assumptions of the scene and surface, and the contributions of each neighbor are assigned on the *basis of heuristics*. In contrast, in *our* approach, which is based on a *supervised learning-based model*, (a) the weights of the network are obtained by training, and (b) after the network has been trained, the weights exhibit an *excitation-inhibition* structure *unlike* any in the literature. The proposed approach imposes no constraint on the disparity gradient in the stereo-pair, is non-iterative, and can be implemented in parallel. The disparity estimates obtained from the proposed approach, as applied to random-dot and natural stereo image-pairs, are comparable to those obtained from the standard cooperative stereo algorithm. However, like the other algorithms of the literature, it does not, *in the present form*, handle occlusion explicitly.

2.1 Proposed Approach

We assume that stereo-pairs with completely known disparities are available for training the **single-layer** network. In fact, we employ **only one stereo-pair with known disparity** for the training phase. As explained in Sec. 2.1.2 later, **this stereo-pair has all range of disparities varying in a periodic manner**. The choice of the single-layer

[1]Hereafter called, for simplicity, the '**network**'.

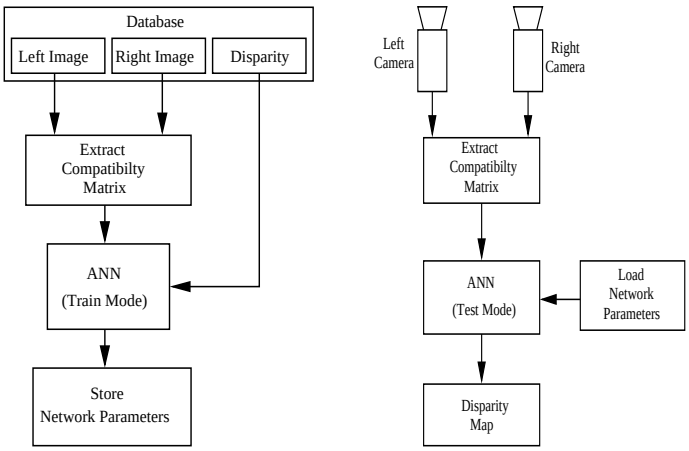


Figure 2.1: *Schematic block diagram of the training and test phase.*

network[2] is motivated by the need to correspond the weights of the network to excitation/inhibition support of the neighboring match values, to help build the model. As a result, the CP is reduced to the problem of system identification.

To train the network (see Fig. 2.3), match values in the neighborhood of each pixel are fed in the form of a matrix, in which each pixel in a segment from the left image (denoted by I_L) is compared with different pixels of the corresponding segment in the right image (denoted by I_R). The resultant matrix is called the **compatibility matrix,** $C_{m,n}(i,j)$, whose argument (i,j) corresponds to the position of the element in the matrix, and subscript (m,n), to the pixel under consideration. We choose a line segment of size, $2D$; and a window of size, $2W$. The size of $C_{m,n}(i,j)$ then is $2D \times 2D$. We assume that disparity ranges from $-D_{max}$ to D_{max}. Therefore, in order that the matrix embeds the correct match within itself, the size of the segment considered must be greater than the maximum disparity that exists in the segment ($D > D_{max}$). In fact, it is fixed at a value higher than the highest disparity value in the image.

Since the match values determined (as in [32]) by individual pixels are vulnerable to noise or distortions in the images, we consider the matches in the neighborhood as well. And, instead of subtracting *merely* the two pixels — one each from left and right images — (i) we choose a small window centered *on* the pixel, (ii) subtract the corresponding pixels, and (iii) sum up the absolute values of the differences in order to get the match value for the two **central** pixels.

It is interesting to note here that some other measures like correlations between the

[2]It is likely that we get better results with multilayer networks but it seems to be difficult to derive the cooperative stereo model from them. This seems to be an interesting open problem.

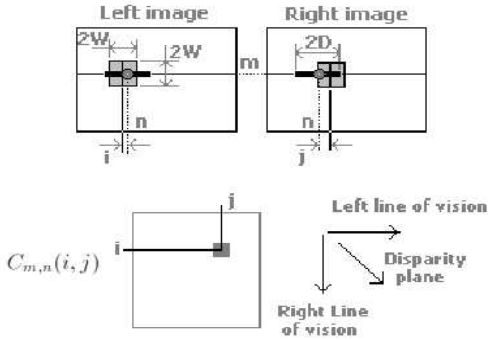


Figure 2.2: *Structure of compatibility matrix*

two windows may also be used in order to improve the results (since the operation involved is a kind of filtering). Evidently, if the initial measures obtained are more reliable, as in *unique matches*, then the network gives better results. This is true with all cooperative stereo models. However, since the quality of the measure also depends on the images, it seems to be difficult to assert the superiority of any measure.

The matrix $C_{m,n}(i,j)$ is computed using a match measure, like correlation or squared differences of windows centered on $(m, n+i)^{th}$ and $(m, n+j)^{th}$ pixels in the left and right images. Therefore, an image of height M and width N will have $M \times N$ number of compatibility matrices. For matching, we use the sum of the absolute pixel-by-pixel difference of the windows centered on the $(m, n+i)^{th}$ and $(m, n+j)^{th}$ pixels in left and right images, respectively (Fig.2.2), and calculate $C_{m,n}(i,j)$:

$$C_{m,n}(i,j) = \sum_{p=-W}^{W} \sum_{q=-W}^{W} |I_L(m+p, n+i+q) - I_R(m+p, n+j+q)|, \tag{2.1}$$

where $i, j = 1, 2, ..., (2D-1), 2D$; $D > D_{max}$. The matrix elements are normalized so as to lie between 0 and 1. Note that a low value of the matrix element implies a better match.

We consider only the interactions along the horizontal direction, i.e., only horizontal line segments are used to construct the compatibility matrix. This procedure can be extended to include vertical neighborhoods in order to get a 3D-compatibility volume [1].[3]

[3]In fact, for general stereo-pairs, it does lead to marginal improvements in disparity but at a price:

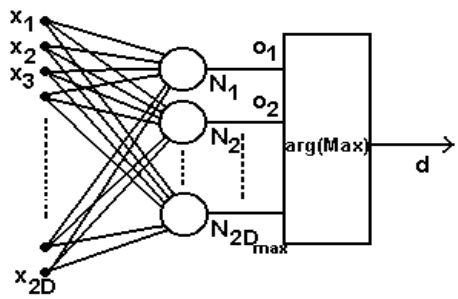


Figure 2.3: *Neural network structure*

In the compatibility matrix, the lines of vision lie along the rows and columns, i.e., $C_{m,n}(i+k,j)$ and $C_{m,n}(i,j+k)$ (see Fig. 2.2), and the disparity plane lies along the direction of the principal diagonal, i.e., $C_{m,n}(i+k,j+k)$. In order to dispense with the problem of finding a correct support region, we consider all the neighboring match points (i.e., the elements of compatibility matrix) without imposing any constraint on them. After the network has converged, the excitation- and inhibition-support regions can be determined, based on thresholding the absolute value of the weights. As a consequence, the network will be able to prune the neighborhood by itself.

In the next subsection, we describe the architecture of the neural network which makes use of the compatibility matrix as an input in order to learn the neighborhood relations.

2.1.1 Neural Architecture

We use a linear neuron model in which each neuron acts as a disparity-selective cell, i.e., for a disparity value, a particular neuron gets excited. Hence the network has $2D_{max}$ number of output nodes, one for each disparity value. The compatibility matrix (of size $2D \times 2D$) is fed to the network as a $4D^2 \times 1$ vector. Note that the network has $4D^2$ input nodes, and is fully connected (Fig. 2.3), i.e., each output node is connected to all the inputs.

Recall that I_L and I_R are the left and right images of a stereo pair. Let I_D be the disparity map. By convention, $I_D(m,n)$ is the disparity value of $I_L(m,n)$ for the case of ideal correspondence. Note that $-D_{max} < I_D(m,n) < D_{max}$. We then have

$$I_L(m,n) = I_R(m, n + I_D(m,n)). \tag{2.2}$$

disparities at the horizontal boundaries of objects are distorted.

The compatibility matrix corresponding to each pixel in the left image is fed as input to the network, and the $(I_D(m,n) + D_{max})^{th}$ output node is forced to fire. Let d_i be the desired response of the ith neuron. In what follows, $i = 1, 2, ..., 2D_{max}$ and $j = 1, 2, ..., 4D^2$. Then

$$d_i = \begin{cases} 1 & : \text{ if } i = I_D(m,n) + D_{max} \\ 0 & : \text{ elsewhere.} \end{cases} \tag{2.3}$$

Let $w_{i,j}$ be the weight connecting j^{th} input node to the ith output node. Since we consider a linear node, the output of the ith node o_i is

$$o_i = \left(\sum_{j=1}^{4D^2} w_{i,j} * x_j \right) \tag{2.4}$$

where x_j is the jth element of the vector constructed from $C_{m,n}$.

With d_i given by (2.3), the quadratic error function, E, is defined as

$$E = \frac{1}{2} \sum_{i=1}^{2D_{max}} (o_i - d_i)^2 \tag{2.5}$$

which is to be minimized by updating the weights of the network. *These weights*, it should be noted, *form the neighborhood supports.* The optimization problem is a standard one. According to the gradient descent strategy, the change in the weight $w_{i,j}$ in the nth iteration is given by,

$$\Delta w_{i,j}^{(n)} = -\eta \frac{\partial E}{\partial w_{i,j}^{(n)}}, \tag{2.6}$$

where η is a pre-defined learning constant and,

$$\frac{\partial E}{\partial w_{i,j}^{(n)}} = (o_i - d_i) \frac{\partial o_i}{\partial w_{i,j}^{(n)}} \tag{2.7}$$

$$\frac{\partial o_i}{\partial w_{i,j}^{(n)}} = x_j. \tag{2.8}$$

The update equation is given by

$$w_{i,j}^{(n+1)} = w_{i,j}^{(n)} + \Delta w_{i,j}^{(n)}. \tag{2.9}$$

The feature vectors derived from the compatibility matrix are fed sequentially to the network which is trained by using (2.9). While testing the network for other stereo-pairs,

the output disparity value is decided on the basis of the winning node, i.e., the node which has the highest output. Therefore, the disparity of the $(m,n)^{th}$ pixel is given by,

$$I_D(m,n) \;=\; arg\max_k(o_k), \;\; for \;\; k = 0, 1, ..., D, \tag{2.10}$$

where (to recall) o is the output of the network, for the input x, constructed from the compatibility matrix $C_{m,n}$.

2.1.2 Network Training and Testing

To train the network, we need synthetic image pairs with known disparity maps. Such pairs are generated by shifting sections of random dots in the left image by the number of pixels given in the disparity map to get the right image. The disparity map should have all the disparity values that the network is supposed to work for, and, to effectively train the network, the disparity values must be presented to the network in a periodic manner, so that the network doesn't enter a plastic state. See Fig. 2.4(a)-(b) for a typical stereo-pair of size 512×512 used for training; and Fig. 2.4(d), for the corresponding true disparity map with disparity values ranging from -5 to +5.

In the course of training the network, we scan image pixels along the rows, and feed in the corresponding compatibility matrices to the network. As a consequence, the disparity values are vertically aligned, and the rows of the disparity map form a periodic pattern.

We test the performance of the network using the following stereo-pairs:

1. Random dot stereo-pairs (RDS) (Fig. 2.8) are synthetic, having randomly distributed dots, whose density controls the amount of texture. They are used to check the performance of algorithms on textured surfaces.

2. Wireframes (Fig. 2.10) contain no clues (like shading or texture) which are important pointers to depth information in (natural) monocular images. They have sparse match-points, with disparity confined to the lines; and (ii) represent the narrow objects (in the image). They help us test the performance of a stereo-algorithm on fine-structured objects.

3. Blurred stereo images (Fig. 2.12) are used to check the robustness of the algorithm. One image in the stereo-pair is blurred using a Gaussian mask.

4. Noisy stereo images (Fig. 2.12) enable us to study the effect of noise on the performance of the algorithm. Noise is added to one image in the stereo-pair.

5. Natural stereo images (Fig. 2.13) contain a combination of textured and texture-less objects.

6. Image pairs with occlusion (Fig. 2.16) act as counter examples, and motivate the development of new algorithms.

7. Stereo-pairs with vertical disparity (Fig. 2.17) are considered to check the robustness of the algorithms with respect to vertical disparities.

Before we present experimental results in the next section, we need to make a few comments on the limitations of the proposed approach (in comparison with those of the literature):

- *If the test stereo-pair has some vertical disparity also, then the proposed approach gives satisfactory results only if the similarity measure is invariant to vertical shifts. With measures like SSD or correlation, the network's performance is reduced with vertical shift of even a single pixel since both SSD and correlation are sensitive to shifts. The same observation holds for the algorithm of Zitnick and Kanade [4] (hereafter called, for convenience, as* **Z-K***).*

  In an attempt to overcome this limitation, the network has been used to extract some appropriate features by feeding in the intensity values directly as input rather than the compatibility matrix. But the network does not seem to converge (even when the network is multilayered).

  Modification of the training datasets (to include those with vertical disparity also) does not help.

- *Perhaps the only way is to start with the assumption that the images also have vertical disparity, and then use two disparity dimensions to find the neighborhood support in the extended space. But even then it cannot handle any scaling or rotation in one image. This is posed as an open problem.*

- *The network can handle* any *disparity gradient it is trained with. But the Z-K algorithm can handle only disparity gradients less than one pixel.*

- *It does not give satisfactory results in the regions of occlusion. (See Sec. 2.2 below.) The same comment applies to the results of [4].*

Table 2.1: *Weight matrix of negative one-disparity node($D = 5$)*

-0.073809	0.020255	0.028104	0.018256	0.021825	-0.029927	0.014415	0.007121	-0.003249	0.002853
0.000780	-0.015041	0.017273	0.020630	0.016672	0.024821	-0.030232	0.003762	0.012003	-0.007972
0.027020	-0.081338	0.018545	0.027727	0.015365	0.019311	0.022052	-0.018890	-0.004954	0.027140
-0.006573	0.047515	-0.125920	0.037975	0.026647	0.009362	0.017251	0.016003	-0.005788	-0.009642
0.001399	0.015326	0.050493	-0.145510	0.047555	0.026938	0.012530	0.021170	0.015555	0.004399
0.023494	0.007472	0.019313	0.048936	-0.153598	0.046643	0.028460	0.014058	0.025717	0.015024
-0.005249	0.023885	0.011610	0.020746	0.038539	-0.155662	0.042567	0.023280	0.016481	0.024902
-0.002315	-0.014045	0.022919	0.023502	0.023045	0.026005	-0.150701	0.034739	0.016650	0.006583
-0.003227	0.001073	-0.026089	0.027589	0.029367	0.021170	0.012056	-0.140953	0.022929	0.006421
0.011311	-0.015714	0.007391	-0.025922	0.030263	0.037956	0.024276	-0.013899	-0.103898	-0.000028

2.2 Experimental Results

See Fig. 2.4 for the training image data in which the left and right image of a gray scale random-dot stereo have the disparity map shown in Fig. 2.4(d), with disparity values ranging from -5 to 5. The network of Fig. 2.3 is trained with a segment size of 20 ($D = 10$) and window size 6 ($W = 3$). The other parameters of training are: size of the network $=$ 11 nodes; and learning parameter, $\eta = 0.1$. The total time required for training the network using the images of Fig. 2.4 (size 512×512) is about 20 minutes, on a P4 $700MHz$ PC, with $256MB$ RAM. However, it is to be noted the training is off-line, and, once the network is trained, disparity computation for a test image pair of the same size is about 2.3 sec.

The network weights for each node are shown in the form of an image in Fig. 2.6. Typical values for a trained network with $D = 5$ are given in Table 2.2.

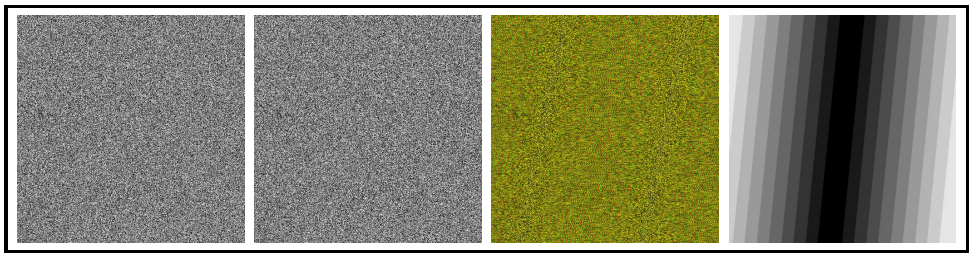
We compare the disparity maps obtained using the proposed approach with those of the Z-K algorithm [4] and the SSD method. A performance measure is defined based on the availability of the true disparity maps. In the case of stereo-pairs with true-disparity map, the percentage of pixels having an error in disparity of more than one pixel, termed *bad pixels*, is used. For a stereo-pair whose disparity map is not available, the right image is reconstructed from the left image using the disparity map, and the difference in the intensities of the (i) reconstructed and (ii) the true right images is used to find the bad pixels.

Example 1: Trained stereo-pair

The weights obtained after training are used to estimate the disparity of the same stereo pair, and the result is compared with the disparity map obtained using the Z-K algorithm in Fig. 2.5. The network estimated disparity has 0.0417% pixels with error greater than one disparity value, and 0.29% for the Z-K algorithm. We find that network performs better than the Z-K algorithm for the images which were used for training. The average response of each node for different disparity values for the stereo pair of Fig. 2.4 is given in Fig. 2.5(a). The plot confirms that the nodes work as disparity-selective cells.

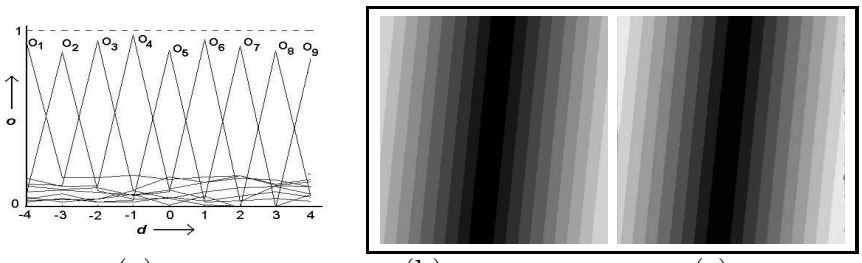
The weights connecting an output node and the elements of the compatibility matrix are displayed in the form of gray-scale image (negative) for all nodes in Fig. 2.6. The images show that excitations are developed in the direction of the principal diagonal, consistent with the continuity assumption; and, as the disparity value increases, the excitation shifts, away from the principal diagonal, to different disparity planes.

Figure 2.7 shows the 3D plot of the weights for the zero-disparity node. Along the principal diagonal, we find a high excitation in a Gaussian-like form, with its maximum at the center of the matrix. But, in contrast with the consequences of the uniqueness assumption, which predicts high inhibition for all other disparity planes, the 3D plot shows high inhibitions only in the neighboring disparity levels (see Fig. 2.7). **Therefore, we arrive at a new model of excitation-inhibition, shown in Fig. 2.7(d), which is distinct from other models of Fig. 1.26**. The new model consists of an excitation for the same disparity plane, with a Gaussian-like tail, and an inhibition for the immediate neighboring disparity planes, with a faster decay.



(a) (b) (c) (d)

Figure 2.4: *(a) and (b) Random-dot stereo pair (c) corresponding anaglyph and (d) its true disparity map ($D_{max} = 5$).*



(a) (b) (c)

Figure 2.5: *(a) Average neuron responses for different disparities, Estimated disparity map using (b) the network (c) Z-K algorithm.*

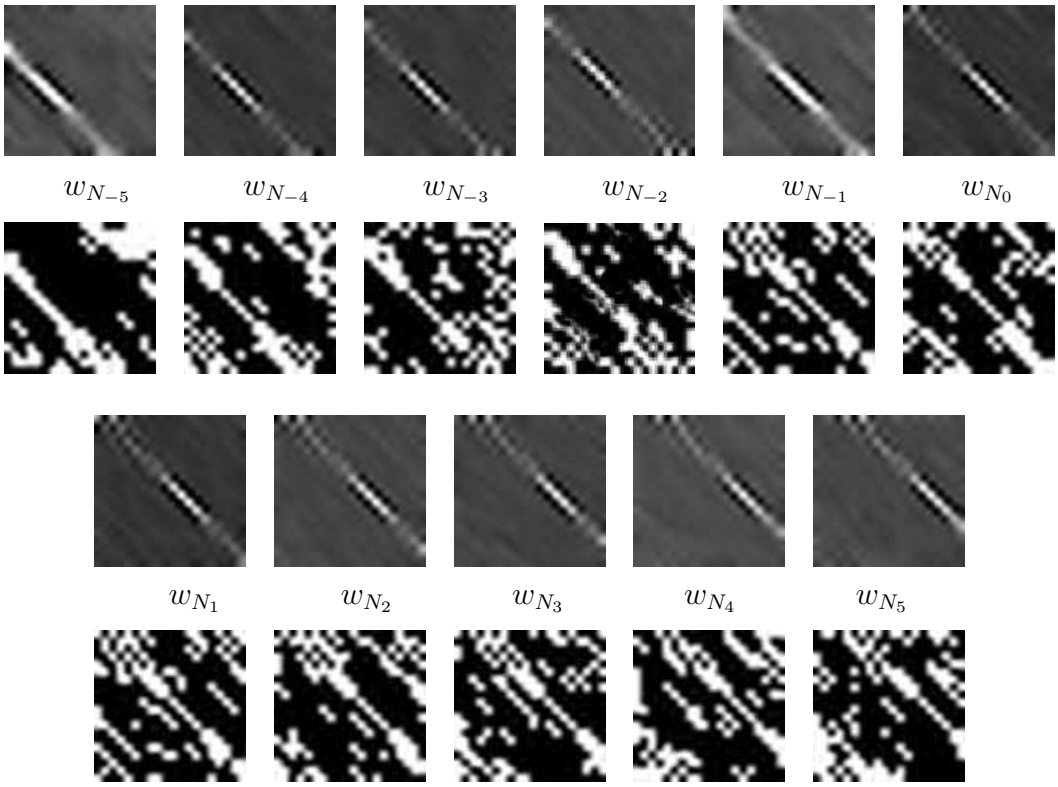
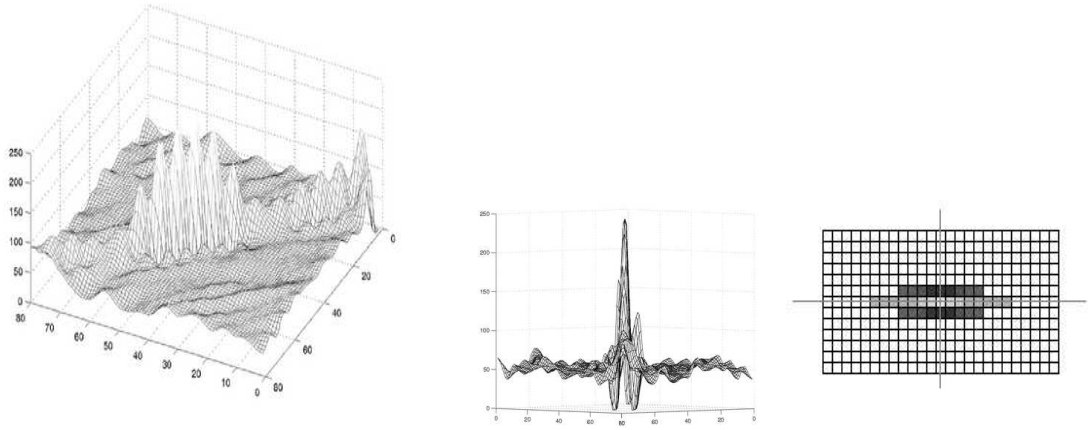


Figure 2.6: *2D image plot (negative) of weights of different disparity nodes, and their corresponding binary excitation (white) - inhibition (black) plots for $D = 10$ and $D_{max} = 5$.*



(a) (b) (c)

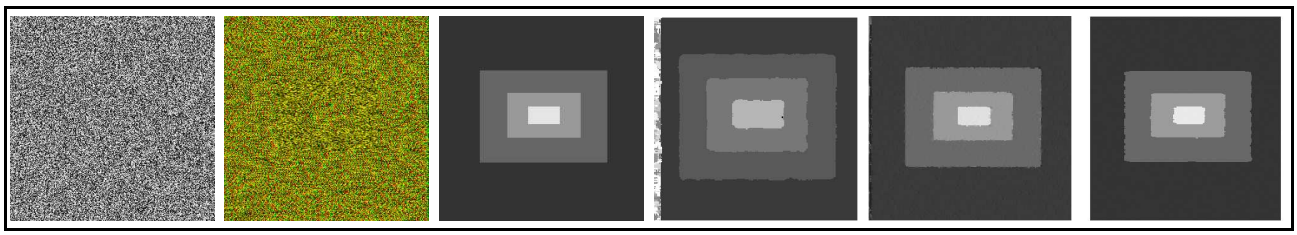
Figure 2.7: *3D plot of the weights of the zero disparity neuron: (a) general view (b) side view along the principal diagonal (c) the* **new Excitation-Inhibition model** *for cooperative stereo. Compare with Fig. 1.26.*

Example 2: Random-dot stereo-pair

For a gray-scale RDS (*test stereo-pair*), with a wedding-cake

like disparity map, the estimated disparity maps using the Z-K algorithm and the network are given in Fig. 2.8; and the corresponding errors in Table 2.2. Similarly, for the binary random dot stereo with 90% sparsity but with the same disparity map, the *estimated* disparity maps are given in Fig. 2.9. The results show that the network is

able to handle sparse textures. But the error measure shows that the performance deteriorates when the test image differs *substantially* from the training image, as in case of binary RDS.

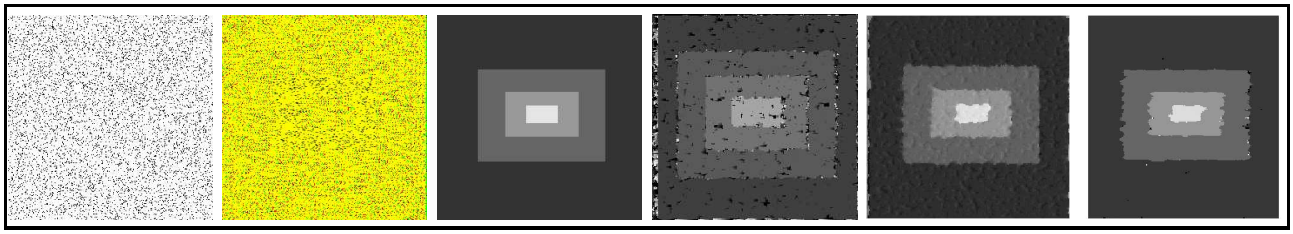


(a) (b) (c) (d) (e) (f)

Figure 2.8: *(a) left image of gray-scale random dot stereo pair (b) corresponding anaglyph; (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network.*

Table 2.2: *Error measure for synthetic stereo-pairs*

Stereo-Pair	SSD	CS	ANN
RDS(Gray)	1.851	0.230	0.186
RDS(Binary)	5.103	1.873	2.988
Wireframe	10.213	67.976	14.873

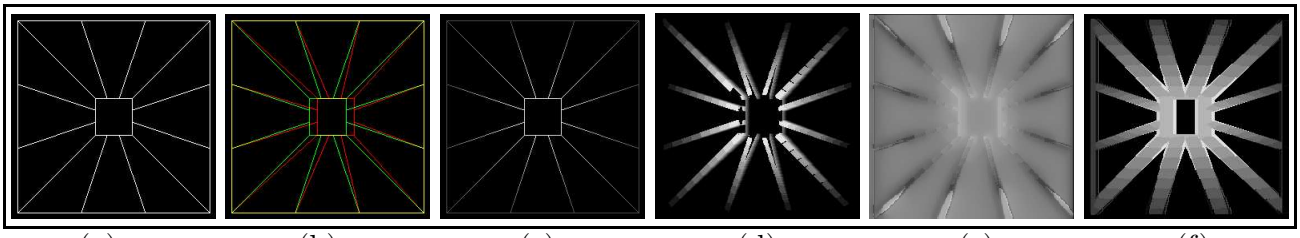


(a) (b) (c) (d) (e) (f)

Figure 2.9: *(a) Left image of binary random-dot stereo (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network.*

Example 3: Wireframe stereo-pair

The estimated disparity map for the wireframe stereo-pair is shown in Fig. 2.10. The network produces a band of disparity instead of a line, as the match measure is based on block matching. But when the estimated disparity map is compared with that of Z-K algorithm, it is found to be 'clean'. In fact, the network is able to estimate correct disparities even at the intersection of lines where *all the other* block matching algorithms fail. Strangely enough, it is not obvious (in the proposed approach) how this happens.[4] When applied to other wireframe stereo-pairs, similar results are obtained.



(a) (b) (c) (d) (e) (f)

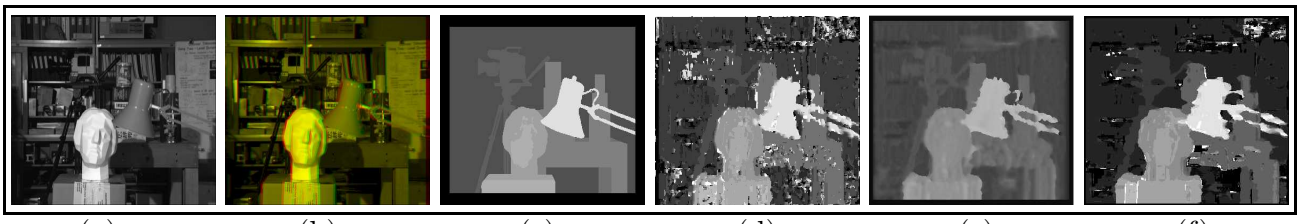
Figure 2.10: *(a) left image of wireframe stereo pair (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network.*

[4]This seems to be an open problem.

Example 4: Tsukuba stereo-pair

See Fig. 2.11 for the stereo-pair of a Tsukuba sequence, with its corresponding anaglyph and the true- and estimated disparity maps. Fig. 2.12 shows two versions of the right image: one that has been subjected to noise (Fig. 2.12(a)) and the other to blur (Fig. 2.12(f)). The corresponding anaglyphs are shown in Fig. 2.12(b) and (g), along with the *estimated* disparity maps. The comparative performance measures for SSD and the Z-K algorithms are given in Table 2.3. It is found that the network *fails* to give correct disparities in textureless regions. In this case, both SSD and Z-K also produce false results.[5] Further, the network result is noisy when compared with that of the Z-K algorithm.

Remark: In textureless regions, all the nodes respond since they see matches at all the disparity values. As a consequence, the disparity map obtained from the network is noisy when compared with that of the Z-K algorithm which is iterative. Therefore, it is *likely* that, if the network is used *iteratively* with an appropriate convergence condition, the disparity map obtained from *it* can be improved. In fact, how to extend the network to accommodate iterative computation is an open problem.



(a) (b) (c) (d) (e) (f)

Figure 2.11: *(a) left image of Tsukuba stereo pair (b) anaglyph (c) true-disparity map; and results obtained from/by (d) SSD (e) Z-K (f) Network.*

Table 2.3: *Error measure for Tsukuba stereo pair.*

Stereo-Pair	SSD	CS	ANN
Tsukuba	8.3659	7.8902	7.9913
Tsukuba(Blur)	22.1079	13.9093	18.2237
Tsukuba(Noise)	36.2641	10.1027	16.8314

[5]However, when the textureless region is small but has a textured background, the Z-K algorithm can produce correct disparity.

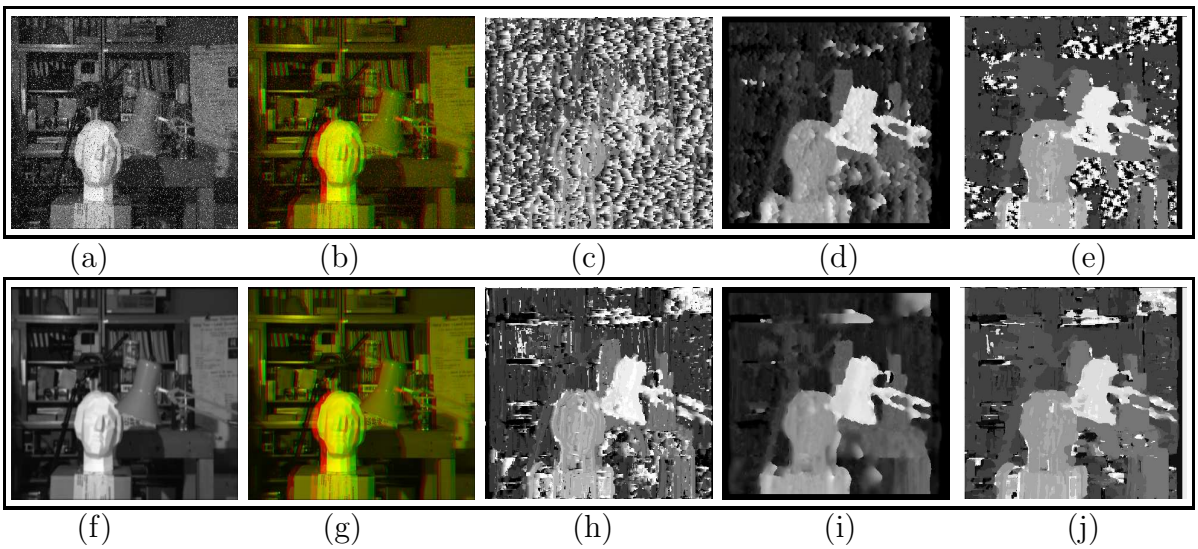


Figure 2.12: *(a), (f) right image with noise and blur respectively, (b), (g) anaglyph; and results obtained from/by (c), (h) SSD, (d), (i) Z-K (e), (j) Network.*

Example 5: Standard stereo-pairs with known disparity map

The disparity maps for some more standard stereo-pairs with known true-disparity maps are shown in Fig. 2.13, and the corresponding performance measures presented in Table 2.4.

Table 2.4: *Error measures for the natural images with true-disparity maps.*

Stereo-Pair	SSD	CS	Network
Venus	24.7182	6.8944	9.3159
Cones	17.8936	11.7814	14.5310

Example 6: Standard stereo-pairs whose true disparity maps are unknown

The results for pentagon and renault stereo-pairs are given in Fig. 2.14, and the corresponding texture-mapped 3D depth plot for pentagon is given in Fig. 2.14. The true disparity maps of both the stereo-pairs are unknown, hence an error measure based on the reconstruction of right image from left is used, and the values are given in Table 2.5. The network gives a good disparity map for the pentagon stereo-pair due to presence of high texture. And, in the case of the renault stereo-pair, the disparity map is as good as that of the Z-K algorithm.

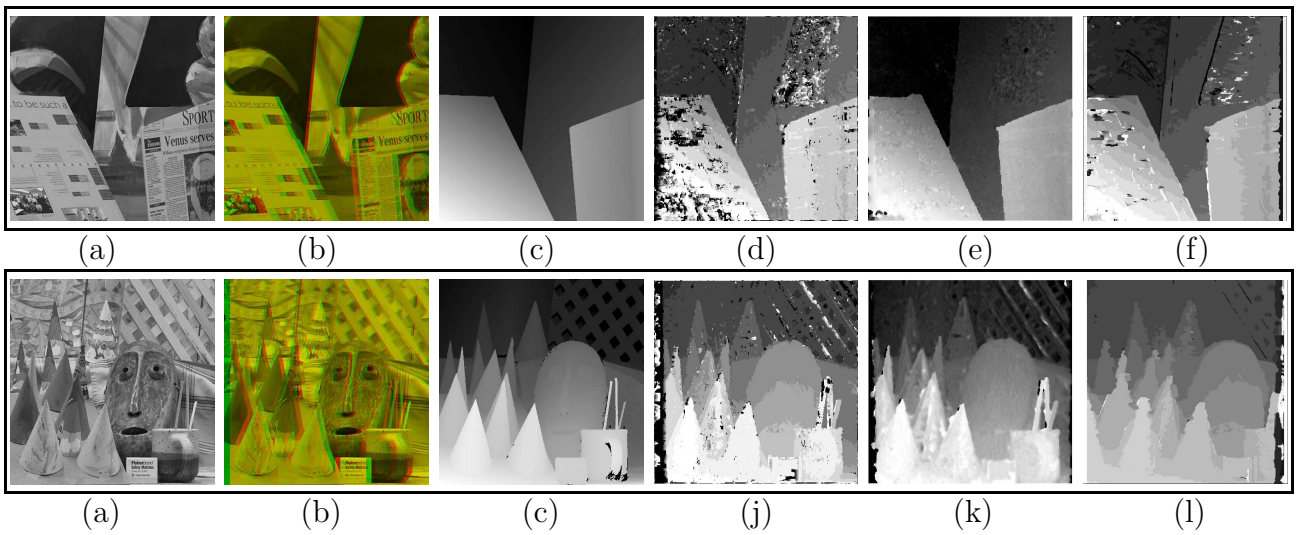


Figure 2.13: *Left image, anaglyph, true-disparity, SSD, Z-K, and Network results for: (a)-(f) Venus stereo pair, (g)-(l) Cones stereo pair, respectively.*

Table 2.5: *Error measures for the natural images with unknown true-disparity maps.*

Stereo-Pair	SSD	CS	Network
Pentagon	14.5190	8.7945	8.1387
Renault	12.4922	6.1850	8.3501

Example 7: Stereo-pair with occlusions

In Fig. 2.16, we consider an RDS stereo-pair with a disparity step of 7 pixels, giving rise to occlusion. It is seen that *all* the algorithms under consideration cannot detect the occluded part. But the (proposed) network and the Z-K algorithms produce a uniform disparity over the occluded region, whereas the SSD produces noisy disparity.

Example 8: Stereo-pairs with vertical disparity

The right image of the Tsukuba stereo-pair is shifted up by 5 pixels to produce the non-epipolar stereo-pair. The results given in Fig. 2.17 show that all the algorithms are badly affected by the vertical shift.

Many more experiments (employing synthetic and natural stereo-image pairs) [39] have been conducted to test the performance of the proposed network-based algorithm, and comparisons made with the results of the literature.

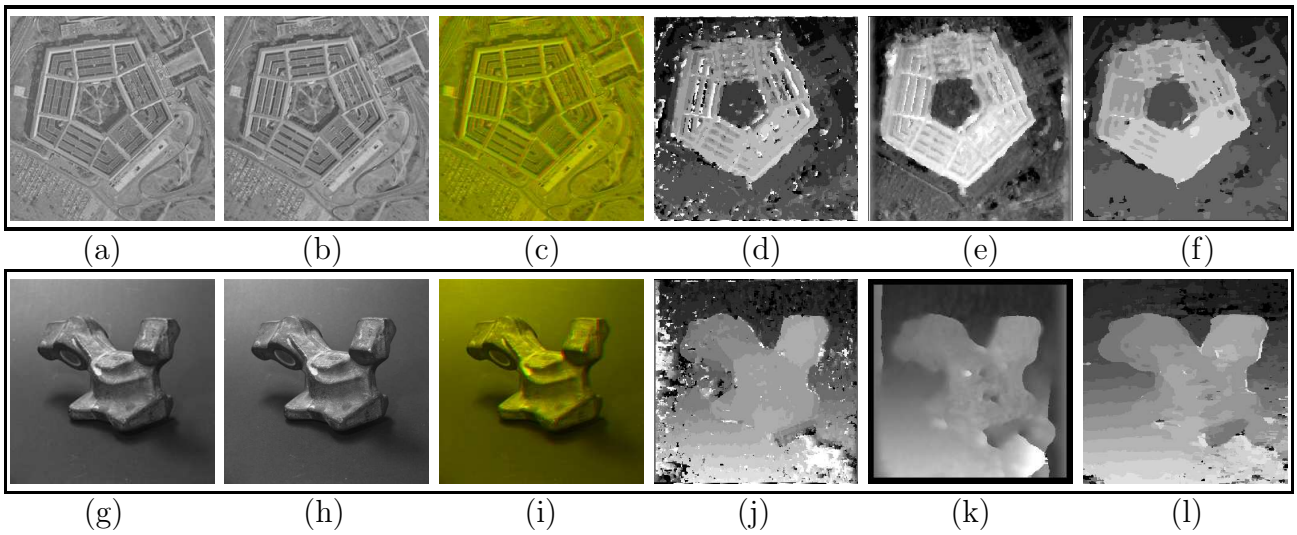


Figure 2.14: *(a), (b) Left and right images of pentagon stereo pair (c) anaglyph (d) estimated disparity map using SSD (e) Z-K algorithm (f) Network, with $D = 10$ and $W = 3$*

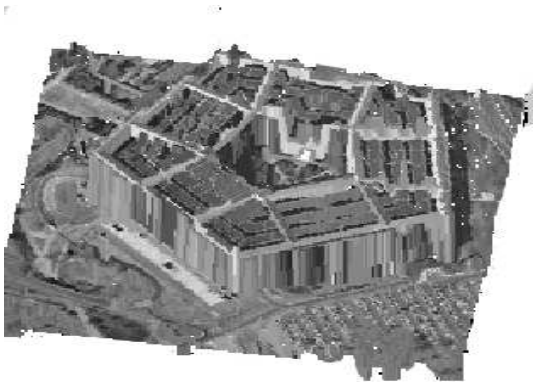


Figure 2.15: *Texture-mapped 3D plot of the estimated disparity map of Fig. 2.14 using the network.*

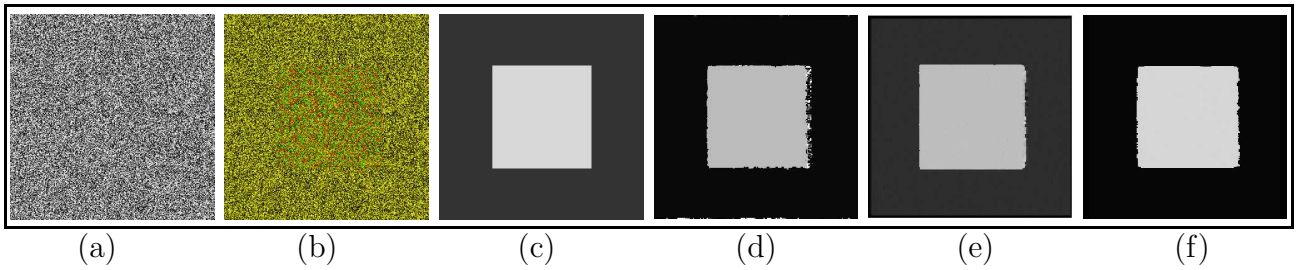


Figure 2.16: *(a) Left image of RDS stereo pair (b) True disparity map (c) anaglyph (d) estimated disparity map using SSD (e) Z-K algorithm (f) Network.*

(a) (b) (c) (d) (e)

Figure 2.17: *(a) Left and right image of Tsukuba stereo-pair with right image vertically shifted by 5 pixels (c) estimated disparity map using SSD (d) Z-K algorithm (e) Network.*

2.3 Summary

Based on the concept of cooperative stereo, we have proposed a new model for stereo disparity estimation, using a single-layer artificial neural network. The main result is that we determine the network weights using *exemplar* synthetic stereo-image data with known disparity maps. The novelty of the approach lies in the fact that the weights characterize the amount of support from each neighbor of the pixel under consideration.

An interesting byproduct of the proposed approach is that the weights of the network after training suggest a new model of neuronal excitation-inhibition (for stereo-image analysis) that is distinct from other models. This new model consists of an excitation for the same disparity plane that has a Gaussian-like tail, and an inhibition for the immediate neighboring disparity planes, with a faster decay.

The performance of the network compares favorably with that of recent algorithms when applied to random-dot and natural stereo image-pairs. However, its superiority consists in (i) imposing *no constraint* on disparity gradient; and (ii) a non-iterative computation that is fast and that can be implemented in parallel since the compatibility matrix for each pixel can be constructed separately. However, in the present form, like most of the other algorithms of the literature, it does not handle occlusion explicitly. And a generalization of the proposed approach to deal with this is a challenging problem.

Chapter 3

Modified Self-Organizing Neural Network Stereo-Disparity Estimation

Outline: *Without assuming epipolar geometry*, we propose a novel application of a modified self-organizing neural network (SONN) to estimate disparity in stereo-pairs. The technique involves the initialization of the network to, *say*, the right image, and then the computation of the amount of deformation required to transform it into the left image. Its unique characteristic is that the deformation *itself* is a measure of disparity. On application to synthetic and natural stereo-pairs (including wire-frames), the technique has been found to perform better than many existing algorithms. Stereo-pairs of images, having (a) salient features like discontinuous depth values at boundaries and surface wrinkles; and (b) occluded and half-occluded regions, are used not only as illustrations of the technique but also as possible *counterexamples* to bring out its limitations, thereby serving as motivation for further work.

3.1 Proposed SONN-based Approach

The self-organizing map (SOM) of Kohonen [40] is a neural network used to transform a higher-dimension feature space to a one- or two-dimensional discrete map in a topologically ordered fashion. The network is trained in an unsupervised manner such that the final weights of the network tend to follow the distribution of the input feature vectors. During the training, the neuron/node whose synaptic weight is closest (in a Euclidean sense) to the input feature vector is declared the winner. The winner's weight vector is updated so that it moves closer to the input vector in the weight space (see (3.1) below). The topological neighboring neurons in the weight space are also updated, in a weighted manner. The updation at n^{th} iteration for the kth-component of weight vector w^n, with input vector α is given by

$$w_k^{n+1} \leftarrow w_k^n + \eta(n)\; g_n(w_k^w - w_k^n)\; (\alpha_k - w_k^n)\,, \tag{3.1}$$

where w^w is the weight vector of the winning node; η is the learning rate, and $g_n(x)$ is the neighborhood-function, defined below, with the parameter $\sigma(n)$ controlling the influence of the winner on its neighbors:

$$g_n(x) = \exp\left(-\frac{x^2}{2\sigma(n)^2}\right). \tag{3.2}$$

This gives rise to a topology-conserving property for the network. If the network is isomorphic to the data set (see Fig. 3.1), i.e., the number of nodes is equal to the number of data points, then, on convergence, each node maps onto a unique data point; and the positions of the nodes represent the best topographic fit for the initial distribution of the nodes.

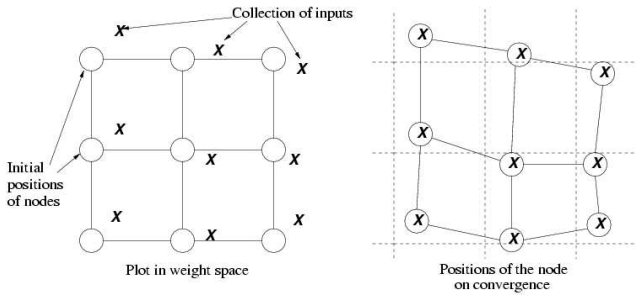


Figure 3.1: *Illustration of topology-conserving isomorphic map using SOFM*

Stereo-disparity estimation, as explained in Sec. 1, involves mapping a pixel in, say, the left image to a "corresponding" pixel in the right image of the stereo-pair. We now present the application of the SOM to discover a possible isomorphic mapping between the two images. It is to be noted here that the standard SOM algorithm *needs to be modified* to take into account the stereo-matching constraints.

The modified SOM algorithm runs as follows:

- Initialize a node for each pixel in the left image, with the corresponding coordinates and intensity as the initial weights. Let $(w_1^{ij}, w_2^{ij}, w_3^{ij})$ denote the weights of the node corresponding to pixel at (i, j) with intensity w_3^{ij}.

- For each pixel in the right image associate a feature vector whose elements are the coordinates and the intensity. These feature vectors serve as inputs to the network.

- Select a pixel at random from the right image, and feed the corresponding feature vector

as input to the network. Let $(\alpha_1^{mn}, \alpha_2^{mn}, \alpha_3^{mn})$ be the input feature vector corresponding to pixel at (m, n).

- Let (p, q) be the index of the winning neuron in the network for $(m, n)^{th}$ input, then

$$(p, q) = \underset{i,j}{\operatorname{argmin}} \sqrt{\sum_{k=1}^{3} (w_k^{ij} - \alpha_k^{mn})^2}$$

The winner node (p, q) corresponds to the minimal distance, and represents the pixel in the left image *which could be a match* for the (m, n)th pixel in the right image.

- Let M and N denote the height and width of an image, respectively. Update *only the first two components* of all the neuron weight vectors as follows:

$$w_k^{ij} \leftarrow w_k^{ij} + h_k(i', j')\, g_k(\Delta I) \left(\alpha_k^{(m+i')(n+j')} - w_k^{ij}\right), \tag{3.3}$$

where

$$\begin{aligned} i' = i - p,\ j' = j - q, \\ h_k(i', j') = \eta_k \exp\left(-\frac{i'^2 + j'^2}{2\sigma_{hk}^2}\right) \end{aligned} \tag{3.4}$$

$$\begin{aligned} g_k(\Delta I) = \exp\left(-(\Delta I)^2 / 2\sigma_{gk}^2\right) \\ \Delta I = (w_3^{pq} - w_3^{ij}), \end{aligned} \tag{3.5}$$

for $k = 1, 2$, and $\forall i, m \in \{1, 2, \ldots, M\}$, $\forall j, n \in \{1, 2, \ldots, N\}$, where η_k is the standard learning rate and σ_{hk}, σ_{gk} are the neighborhood parameters, which control the rate of propagation of disparity in the topological neighborhood and within the range of the object, respectively.

Repeat the above three steps N_p times where N_p is a pre-determined number. In our experiments, we have used $N_p = 100 \times MN$.

- The disparity vector (d_x^{ij}, d_y^{ij}) for each pixel (i, j) is defined as:

$$d_y^{ij} = i - w_1^{ij} \tag{3.6}$$

$$d_x^{ij} = j - w_2^{ij} \tag{3.7}$$

Here (3.6) and (3.7) represent the vertical and horizontal disparities, respectively.

We call the above algorithm the modified self-organizing feature map, SOFM (or **MSOFM**). The updation mechanism in MSOFM is believed to be a significant modification of the standard SOM as explained below in Sec. 3.2. A simplified block diagram of the above mentioned approach is shown in Fig. 3.2, where N is the number of iterations. Note that MSOFM assumes that the mapping of each pixel from one image to the other is isomorphic. But, in practice, some sections of one image will be occluded in the other. For instance, Fig. 3.10 shows a left image of a stereo pair and the actual disparity map. Assuming isomorphism, the above algorithm yields a distorted disparity map (see Fig. 3.10 ($\mathcal{F}_{RL}$)). For such images, the MSOFM needs to be further generalized.

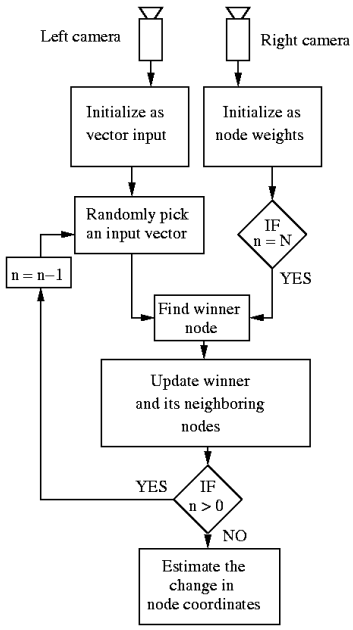


Figure 3.2: *Block diagram of the MSOFM approach.*

3.2 Analysis of MSOFM

The MSOFM converts the stereo-correspondence problem into an estimation of an *onto* map from each pixel of one image to the *corresponding* pixel in the other image of the stereo-pair.

Let S_L represent the set of pixels in the left image, and S_R, the set of pixels in the right image; $S_x = \{[x_x, y_x]^t\}$, the set of position vector of the pixels; and $I_x(s)$, for $s \in S_x$, is its intensity. Then the correspondence map is represented as the transformation, $\mathcal{F}_{RL} : S_R \rightarrow S_L$ (see Fig. 3.3), which is estimated based on the following constraints:

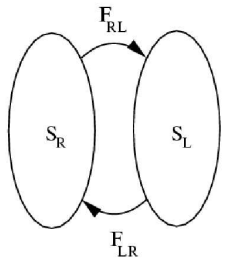


Figure 3.3: *Mapping process.*

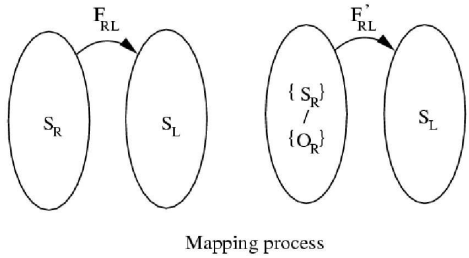


Figure 3.4: *Mapping process before and after occlusion analysis.*

1. *Photometric constraint:*

$$Minimize \sum_{s \in S_R} |I_R(s) - I_L(\mathcal{F}_{RL}(s))| \tag{3.8}$$

This is a standard constraint used in all correspondence algorithms, according to which the corresponding points should have *similar* intensities. In MSOFM, for $[i,j]^t \in S_R$, the correspondence point is given by $\mathcal{F}_{RL}([i,j]^t) = [p,q]^t$, $[p,q]^t \in S_L$ which is the winner node (see (3.3)). The updation of the winner node in (3.3) tends to reduce the distance between the correspondence-node and the input in the coordinate space, thereby *strengthening* the map.

2. *Continuity:* The transformation should be one-to-one and topologically ordered. We incorporate continuity using,

$$\begin{aligned} ||\mathcal{F}_{RL}(s^1) - \mathcal{F}_{RL}(s^2)|| < \epsilon \\ \text{if } \; ||s^1 - s^2|| < \delta_1 \; \text{ and; } \; |I_L(\mathcal{F}_{RL}(s^1)) - I_L(\mathcal{F}_{RL}(s^2))| < \delta_2 \end{aligned} \tag{3.9}$$

where $s^1, s^2 \in S_R$ and $\epsilon, \delta_1, \delta_2 > 0$.

In contrast with the algorithms of the literature, which define continuity in the coordinate space alone, we use continuity in 3D feature space, with intensity as one of the components. This helps preserve the *discontinuity in disparity* based on the assumption that it is *accompanied by a discontinuity in intensity.*

The MSOFM implements this criterion using (3.3). The terms g and h define the neighborhood in the intensity and coordinate spaces respectively. The schematic diagram in Fig. 3.5 shows the difference in the directions of updation in the SOM and MSOFM. The latter produces a one-to-one map that is also topology-ordering.

The modified updation equation (3.3) ensures that the neighboring nodes take on similar disparity, and hence improve continuity in disparity. For the significance of the term g in updation, see Fig. 3.6.

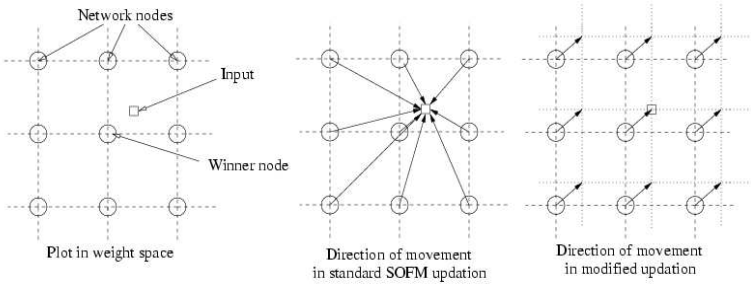


Figure 3.5: *Schematic illustration of the updation process.*

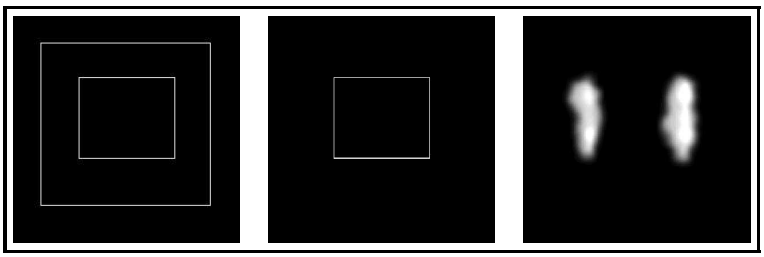


Figure 3.6: *(Left to right) Left image of a wire-frame stereo, corresponding true disparity map and the estimated disparity map without using the g term*

In the standard SOM updation process, the winner and all its neighboring nodes move towards the input vector in the weight/feature vector space. But, for the correspondence problem, we require an isomorphic map. This inconsistency is demonstrated in the case of zero disparity (see Fig. 3.7). Hence, the updation equation is modified to (3), so that the winner node moves towards the input vector, *whereas the neighbors move to corresponding neighbors of the input vector.*

Using the modified updation rule (3), the network gives correct disparity for the zero disparity case. Fig. 3.5 shows a schematic updation process as in the standard SOM and according to (3).

It is important to note here that, in contrast with the application (standard) SOM, the intensity feature, *in MSOFM*, is **not** updated, since such an operation would change the photometric constraint, and produce an erroneous disparity (see Fig.3.8).

3. *Localization:* For stereo-pairs containing periodic textures, if the periodicity of the texture is n pixels, then there is a match at every n pixel interval. If d is the actual disparity at that point, observations suggest that the perceived depth corresponds

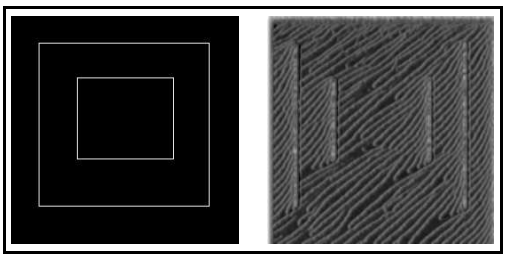


Figure 3.7: *(Left) Image used for both left and right image of stereo pair, i.e. disparity is zero everywhere, (Right) disparity map obtained using standard SOM updation.*

Figure 3.8: *(Left to right) left image, right image of the pentagon stereo pair, intensity weight of the network (right image) after 30 iterations with intensity updation,* and *estimated disparity map.*

to a disparity of min $(d \ mod \ n, n - d \ mod \ n)$, This shows that, out of all the maps possible, we need to select the one that gives minimum disparity. In the solution space given by the above constraints, we minimize $\sum_{s \in S_R} ||s - \mathcal{F}_{RL}(s)||$, thereby ensuring that, in the case of non-unique solutions, we choose the localized solution in the coordinate space. Note that in the case of other algorithms, the *range of disparity has to be known* a priori in order to get a unique solution. In contrast, MSOFM needs no such prior knowledge of disparity range.

4. *Occlusion:* The constraints mentioned above assume no occlusions, which implies that the occluded pixels have to be pre-determined. Below (in Sec. 3.2.1), we describe how the MSOFM inherently detects occluded pixels, and uses a modified pixel set S'_R and S'_L to estimate disparity.

3.2.1 Occlusion analysis

If some point in the scene is not visible to a camera, then the point is said to be *occluded* in the image. In a stereo-pair of images, if an occluded point in one image is visible in the other, then the point is said to be *half-occluded*. The pixel corresponding to this half-occluded point does not have a correspondence-point in the other image. A standard (but implicit) assumption in many stereo algorithms is that *there exists a correspondence*

point for every pixel of one image in the other. In stereo-pairs with half-occluded points, a false mapping is invariably produced with respect to such points. Various pre- and post-processing approaches have been proposed to find the occlusion points as explained in Chapter 1.

Based on the intensity variations in the occluded region (i.e., pixels in the background or of the occluded object having no correspondence points), we classify the occlusions into four categories:

1. Occluded region having an intensity similar to that of the occluding and the occluded objects.

2. Occluded region having an intensity similar to that of the occluding object but different from that of the occluded object.

3. Occluded region having a different intensity from that of both the occluding and the occluded objects.

4. Occluded region having an intensity similar to that of the occluded object but different from that of the occluding object.

Here the similarity is with respect to intensity difference and also the neighborhood. Even though the first three cases occur in general, they lead to ambiguity even for human perception. The fourth case is more common as the occluded region usually belongs to the occluded object itself.

Consider a textureless rectangular fronto-planar object, in a textureless background (Fig. 3.9). The region O_L is occluded in the left image, and the region O_R in the right. Now if the left image is initialized as the network, and the right image as the input, Fig. 3.9 shows the mapping produced by MSOFM for a scan line. The nodes corresponding to O_L will never win in any iteration. Hence the nodes that do not win in any iteration can be classified as the occluded pixels belonging to the fourth category. Similarly, O_R can be determined during the estimation of $\mathcal{F}_{LR}$. Therefore, discovering occlusion is inherent in the mapping process.

Once O_L is determined in the estimation of $\mathcal{F}_{RL}$, a new map $\mathcal{F}'_{LR} : \{S_L - O_L\} \rightarrow S_R$ can be defined, which would be free of false mapping due to occluded points. The mapping processes before and after occlusion detection are shown in Fig. 3.4. Hence during estimation of $\mathcal{F}'_{LR}$, O_L's are not used to produce the map. But the neighborhood updation propagates the disparity value of the occluded object into the occluded region. The same technique works for the third category of occlusion, but the disparity of the occluded region will be assigned to be that of either occluded or occluding object, whichever has

the intensity closer to the occluded region. During estimation of $\mathcal{F}'_{LR}$, it is initialized as $\mathcal{F}^{-1}_{RL}$. This is achieved by initializing the right image as the network, with the coordinate weights given by,

$$w_1^{ij} = i - d_y^{pq} \tag{3.10}$$

$$w_2^{ij} = j - d_x^{pq} \tag{3.11}$$

where $[p, q]^t = \mathcal{F}_{RL}([i, j]^t)$ and d_x, d_y are defined in (3.6) and (3.7). Hence $\mathcal{F}_{LR}([p, q]^t) = [i, j]^t$.

Fig. 3.10 shows the left image of a textureless rectangular plane floating in space. The second image is the *true-occlusion map*, a binary image, in which occluded pixels, O_L, are in white. Note that any pixel which corresponds to uncorrelated noise in the images is also considered to be an occluded point.

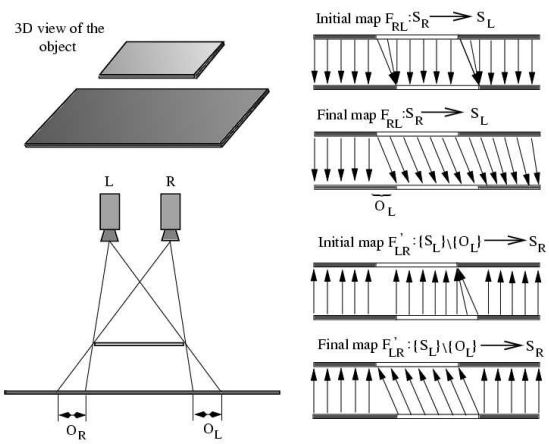


Figure 3.9: *(Top-left) 3D view of the object under consideration, (Bottom-left) Stereo imaging of the object, (Right) Forward and reverse mapping process during disparity estimation.*

3.3 Implementation and Results

We experiment with the application of MSOFM to standard, non-epipolar and noisy stereo-pairs. We compare the obtained results with that of sum-of-squared difference (SSD), cooperative stereo (CS)[4] and graph-cut-based (GC)[20, 21] approaches. From the computational point of view, in order to increase the processing speed, the winner for each

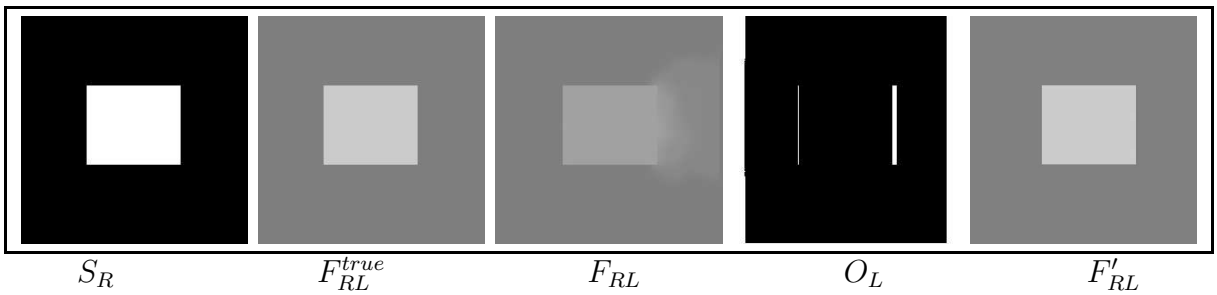


Figure 3.10: *Occlusion analysis of a fronto-planar textureless object. (Left to right) (S_R) right image, (F_{RL}^{true}) true disparity map, (F_{RL}) disparity map estimated by SOM without occlusion analysis, (O_L) occlusion map, (F'_{RL}) estimated disparity map after occlusion analysis.*

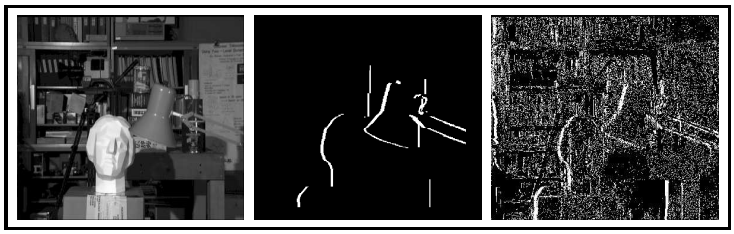


Figure 3.11: *(Left to right) Left image of tsukuba stereo-pair, corresponding true-occlusion map (occluded pixels are in white) and the estimated occlusion map.*

input is determined in a small neighborhood as against searching the whole image; and the updation of weights is carried out within the regions where the Gaussian tails tapers off to 0.001. Typical values of different parameters that are used for the experiments are as follows: (1) number of iterations, $N_p = 100 \times MN$, depending on the size of the image and of the textureless region; (2) neighborhood parameter of h, $\sigma_h = 10.0$ (which controls the spread in disparity); (3) neighborhood parameter of g, $\sigma_g = 10.0$, depending on the intensity levels of the image; and (4) learning constant, $\eta = 0.1$ (which controls the speed of convergence and the smoothness of the disparity map). These parameters are maintained at a constant value, and the network stabilizes as a consequence of the implemented constraints. As far the execution of the MSOFM-based algorithm is concerned, for an image-pair of size 256×256 with a search region of 20×10 pixels, the time taken to compute the disparity map on a Linux platform with Pentium-IV 1.4GHz CPU and 256MB RAM is, typically, 120 secs.

We now present the results obtained on some elementary and standard stereo-pairs including those of non-epipolar geometry and noisy images.

3.3.1 Performance measures for comparison

To compare the results of different algorithms, we adopt the method similar to that of Scharstein *et al.* [27]. In the case of stereo-pairs with true disparity map, the estimated and the true disparity maps are normalized, such that the pixels with lowest disparity are assigned zero value, and the highest disparity value is equal to the range of disparity. Then the difference in the disparity value is computed at each pixel between the estimated and the true value. If the difference is greater than a threshold, then the corresponding pixels are classified as "bad". In our experiments, we have used a threshold of 2. The percentage of bad pixels in the image is used as the measure for comparison, and we call it **percentage of bad pixels in disparity** (PBD).

On the other hand, in the case of stereo-pairs without the true disparity maps, the estimated disparity map is used to estimate the right image from the left image. Then the non-occluded pixels which have difference in intensities between the actual right image and the estimated right image greater than a threshold are classified as bad pixels. The percent thus obtained is called **percentage of bad pixel in intensity** (PBI). A threshold of 50 was used in our experiments. It should be noted that this approach fails to give reliable measures in textureless regions and near the object boundaries.

3.3.2 Elementary Stereo-pairs

See Fig. 3.12 for the results obtained for some combinations of textured and textureless stereo-pairs. Results show that, in the case of textured images, if the intensity distributions in the object and in the background are similar, then the disparity estimate by MSOFM at the boundaries is more noisy than that obtained by other algorithms (Fig. 3.12). On the other hand, in the case of textureless images, SSD is unable to propagate disparity, in comparison with other algorithms. The MSOFM is able to estimate disparity in the case of textureless object and background, whereas the other algorithms fail. In Fig. 3.12 (4th row), the FPS object has a periodic texture with periodicity of 15 pixels, and the object has a disparity of 10 pixels. Observation shows that the object is perceived as having a disparity of -5 pixels. The MSOFM is able to compute the disparity of -5, because of the localization constraint, whereas, with other algorithms, the computed disparity depends on the search range.

If a binary random-dot stereogram is treated as a texture, the MSOFM fails to give correct disparity. This is due to the lack of features for matching but can be resolved by blurring the images using a Gaussian mask so as to introduce some spatially dependent intensity. The disparity results with and without blur are shown in Fig. 3.13.

Even in the absence of features, *like edges or zero-crossings*, the network produces an

Table 3.1: *PBD for elementary stereo-pairs.*

Stereo-Pair [1]	SSD	CS	GC	SOFM
FPS(OB)	1.5575	1.4400	1.5443	2.9843
NFPS($O\bar{B}$)	8.2760	2.4000	8.7576	2.3107
FPS($\bar{O}\bar{B}$)	27.3215	22.4510	9.3866	4.3681
FPS($O'B$)	1.2109	2.0657	1.8008	9.7474
FPS(OB)$_{blur}$	1.5187	1.7578	1.4731	3.1647
WFS	10.2126	67.9764	7.9590	0.8209
LCS	12.7006	22.6512	18.3457	11.2762

Table 3.2: *PBD for the natural images with true-disparity maps.*

Stereo-Pair	SSD	CS	GC	SOFM
Tsukuba	8.3659	7.8902	5.4941	6.5701
Venus	24.7182	6.8944	5.6629	8.2961
Cones	17.8936	11.7814	10.9997	11.1556
Teddy	14.8341	16.0013	24.0699	12.1582

acceptable disparity map (Fig. 3.14). The disparity map produced by SOFM is smooth compared to other approaches. The superiority of the SOFM algorithm is demonstrated in the case of wire-frames, for which the computed disparity is *indeed* confined to the lines, and the disparity map is close to the true-disparity. In contrast, the other algorithms give rise to *blocky* artifacts in the disparity map.

3.3.3 Standard Stereo-pairs

Figure 3.15 shows estimated disparity maps for some standard stereo-pairs. Qualitative comparison of the results shows that MSOFM has the following distinct advantages: (i) It can estimate disparity of fine objects. (ii) It can detect discontinuity in disparity effectively. (iii) It can preserve the shape of objects. And (iv) it can compute a smooth disparity map. The PBD and PBI's for the stereo-pairs are tabulated in Table 3.2 and 3.3.

(1) Stereo-pairs with known disparity map: (a) **Tsukuba:** The stereo-pair has only fronto-planar surfaces. The background is textured with small textureless patches in the foreground. Compared to other algorithms the MSOFM preserves the shape of the objects. But it fails to estimate the disparity of the lamp post. The disparity values close

Table 3.3: *PBI for the natural images with unknown true-disparity maps.*

Stereo-Pair	SSD	CS	GC	SOFM
Pentagon	14.5190	8.7945	11.9412	10.3411
Renault	12.4922	6.1850	9.8001	7.0193

Table 3.4: *PBD for the stereo-pairs with noise.*

Stereo-Pair	SSD	CS	GC	SOFM
Vertical shift	48.3688	28.6301	24.0831	7.4517
Vertical scale	54.0210	41.8365	25.0172	8.3523
Impulse noise	36.2641	10.1027	15.9569	17.8123
Blur	22.1079	13.9093	16.3475	17.1223
Contrast	27.4351	23.5982	71.3943	26.3433

to the outer boundary of the lamp and the camera are noisy, since the occluded parts have similar intensities. (b) **Venus:** It has two textured slant surfaces in the foreground and two partially textured surfaces in the background. The surfaces in the background form a crease. Once again the MSOFM preserves the shapes as also discontinuities in the disparity. (c) **Cones:** The background and the foreground are well textured. The sticks in the mug stand out. The mask has a gradient in disparity accompanied by texture. The MSOFM is able to give correct disparity for the sticks. The graph-cut-based approach is unable to preserve the structures, but overall from the PBD point of view gives a comparable result. (c) **Teddy:** The roof and chimney of the house form a textureless plane surface. The teddy on the floor has good intensity gradient. Only MSOFM is able to estimate smooth disparities in the roof of the house. But it fails to give correct disparity near the right arm of the teddy-bear (close to the top) as the background has similar intensity.

(2) Stereo-pairs with unknown disparity map: (a) **Pentagon:** The image is highly textured. The MSOFM is able to give the structure of the pentagon effectively. It is also able to distinguish the bridge from the ground. The PBI's for different algorithms are given in Table 3.3. (b) **Renault:** The smooth gradient of the ground is accompanied by a gradient in intensity. The fine structures of the machine part are effectively estimated by the MSOFM.

Occluding Contours : Observations by Nakayama *et al.* [41, 42] show that the occluded contours or sections always take on such disparity values as to suggest that they are behind the occluding object. This is clearly visible in the Panum's case (Fig. 1.30), where in reversal of the stereo-pair will change the contours that are matched. None of the existing algorithms take this into account.

Orientation Discontinuities : The SOFM can produce discontinuities in orientation only if there is a change in intensity. In certain cases, as in Fig. 3.17, there are 2 possible disparity maps. One in which only the middle black line appears to be raised above the background. The other in which the white plane forms a roof like structure. Our perception gives rise to the roof-like effect, but the MSOFM gives the disparity map in which only the line is raised. In Fig. 3.15 (venus), the MSOFM is able to estimate disparity correctly in the presence of discontinuity in orientation, accompanied by texture or intensity gradient (see Fig. 3.18).

The slant surfaces can be classified into 2 categories (see Fig. 3.19): (1) Vertical slant: These are plane surfaces which have a rotation about the x-axis. In this case, the uniqueness of the match is maintained. The disparity map has gradient only along the vertical direction. This can be subdivided into two classes based on the texture. (a) Textured: If the texture is dense, most of the algorithms produce an acceptable disparity map. (b) Textureless (Fig. 3.19, 1st row): The disparity map produced depends on the implementation of continuity and uniqueness constraints. Here each pixel within the object has a unique match. But the gradient of disparity in the vertical direction is controlled by the implementation of continuity constraint in the vertical direction.

(2) Horizontal slant: (a) Textured slant surface: If the surface has a dense texture, then the ability of the algorithm depends on the interplay between continuity and uniqueness constraints. (b) Textureless slant surface with boundaries along epipolar lines (Fig. 3.19, 2nd row): The mapping within the object is non-unique. Only the vertical boundaries help estimation of disparity and propagation. Moreover, the disparity has to be propagated from one end to the other. (c) Textureless slant surface, boundaries not along epipolar lines (Fig. 1.27): This is very similar to the above, but the disparities, which can be determined at non-vertical boundaries, help in the propagation of disparity.

Our algorithm, in which the planes are forced to take on fronto-planar form, performs well on textured slant surfaces. But it fails to produce gradient in textureless surface unless there is good support from the boundaries (as in the last case). The results for repeated textures are given in Fig. 3.12 (4th row).

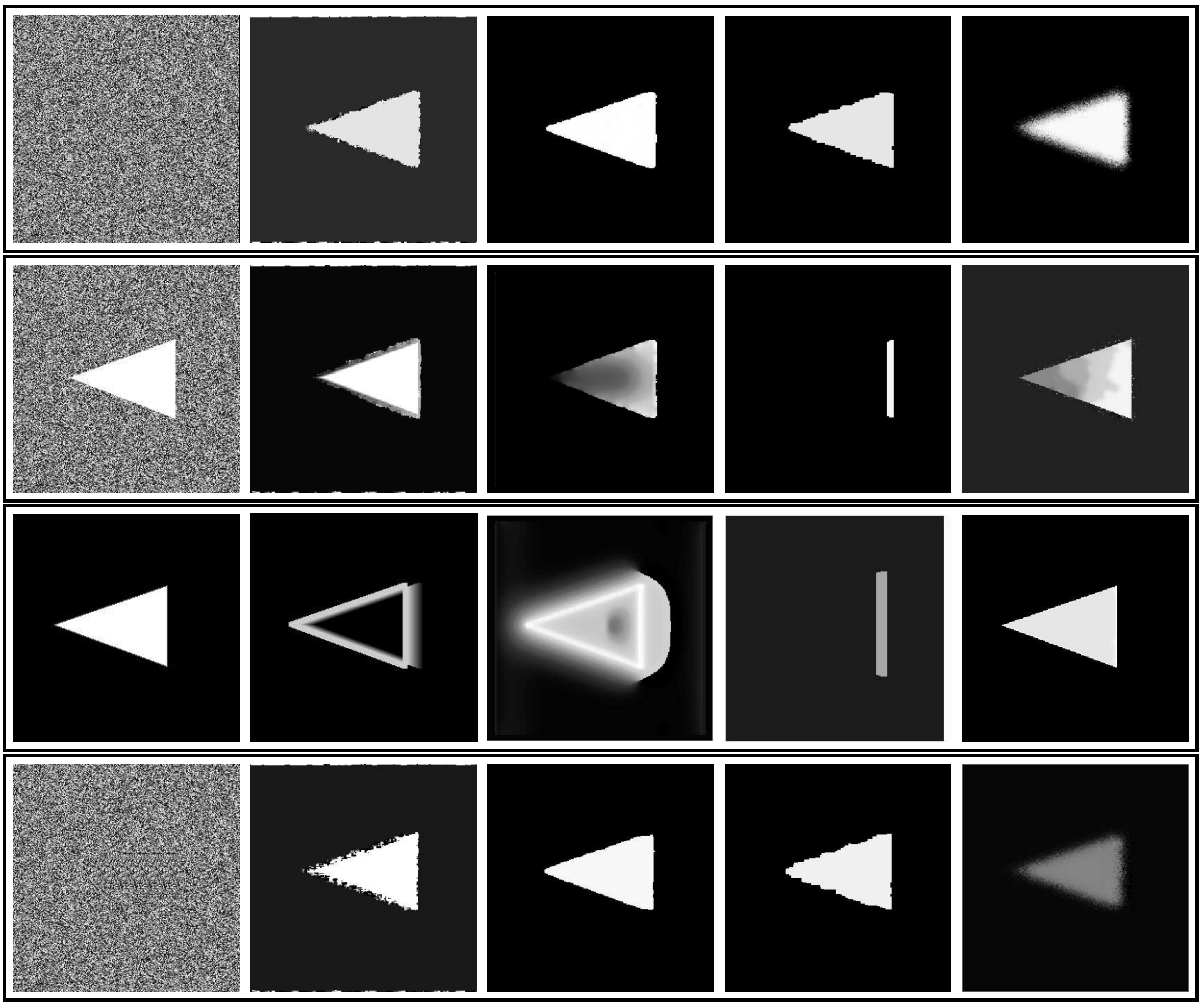


Figure 3.12: *(Left to right) Left image and the corresponding estimated disparity map using SSD, cooperative stereo, Graph-cut, SOFM for (1st row) FPS with textured object and background, (2nd row) NFPS with textureless object and textured background, (3rd row) FPS with textureless object and background, (4th row) FPS with textured object (periodic texture) and background.*

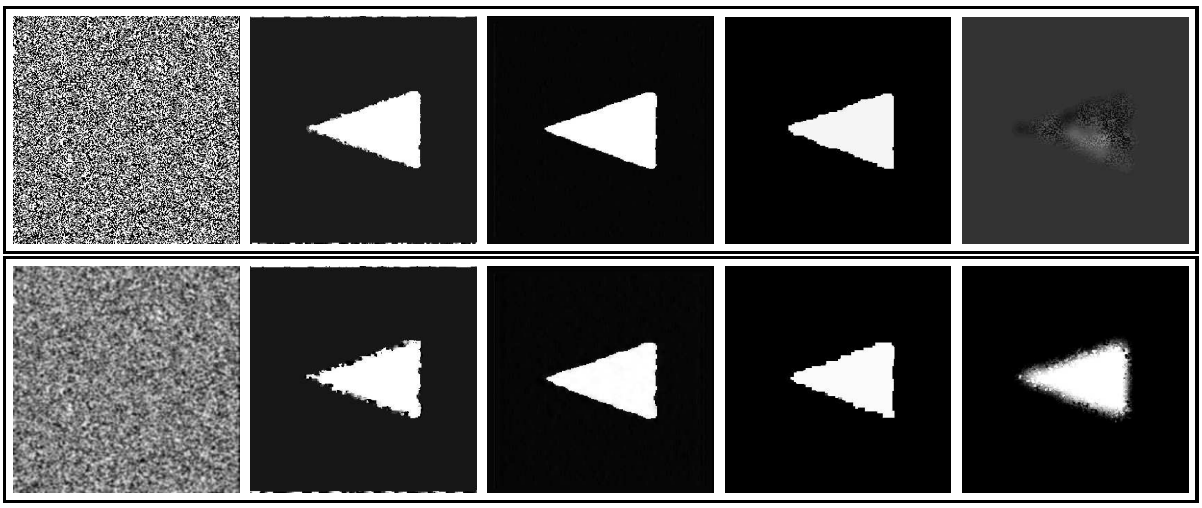


Figure 3.13: *(Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) FPS with binary texture-mapped object and background, (2nd row) Gaussian blurred pair in row 1.*

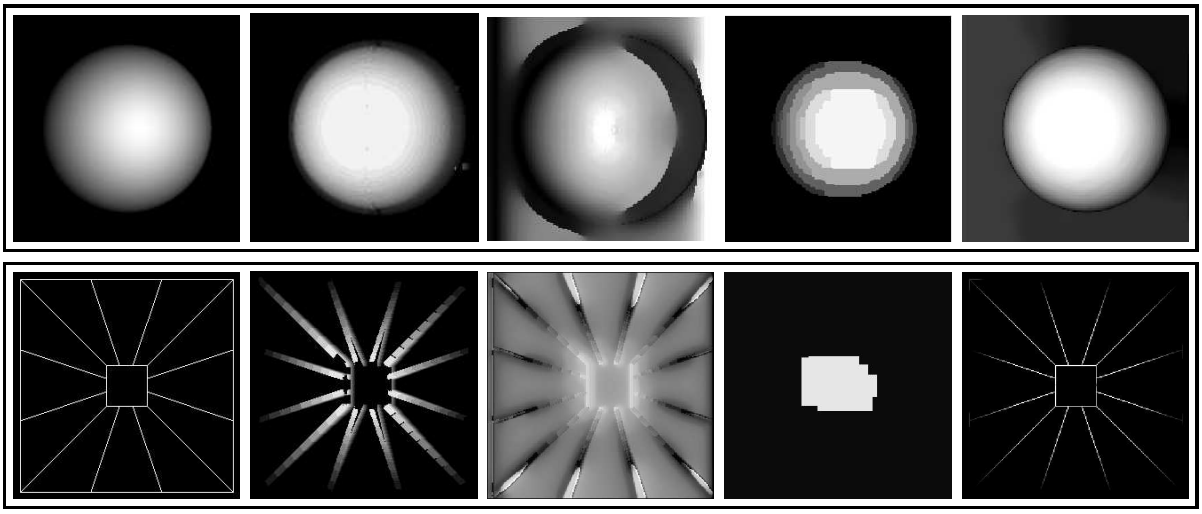


Figure 3.14: *(Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) LCS, (2nd row) WFS.*

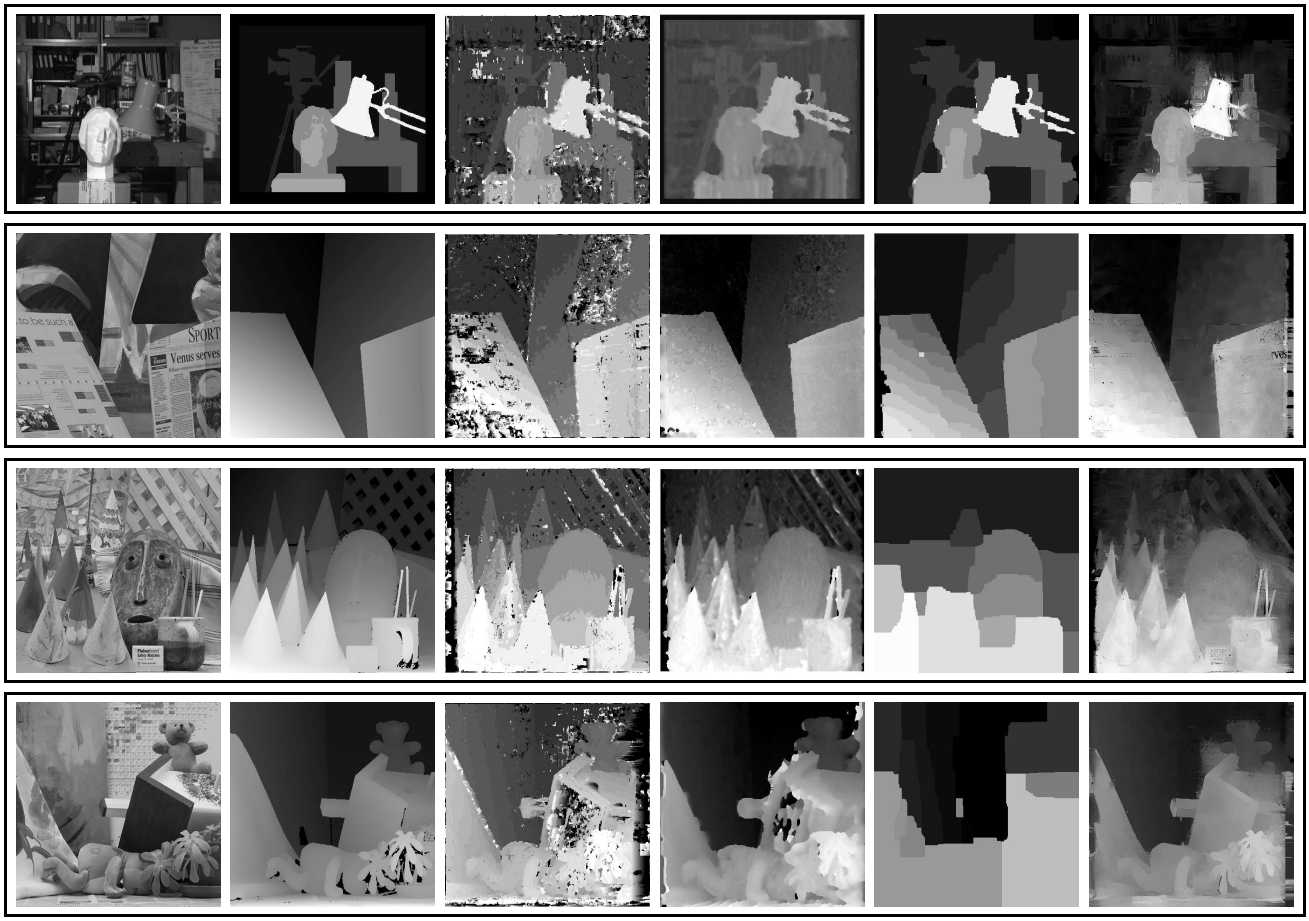


Figure 3.15: *Natural stereo-pairs with known disparity maps: (Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) tsukuba, (2nd row) venus, (3rd row) cones, (4th row) teddy stereo-pairs.*

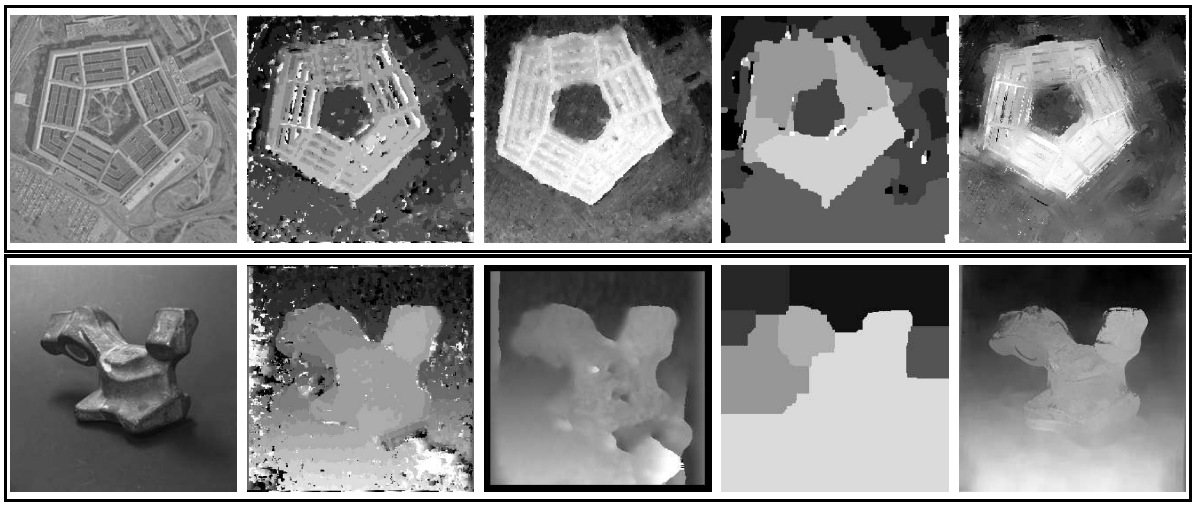


Figure 3.16: *Natural stereo-pairs with unknown disparity maps:(Left to right) Left image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) pentagon, (2nd row) renault stereo-pairs.*

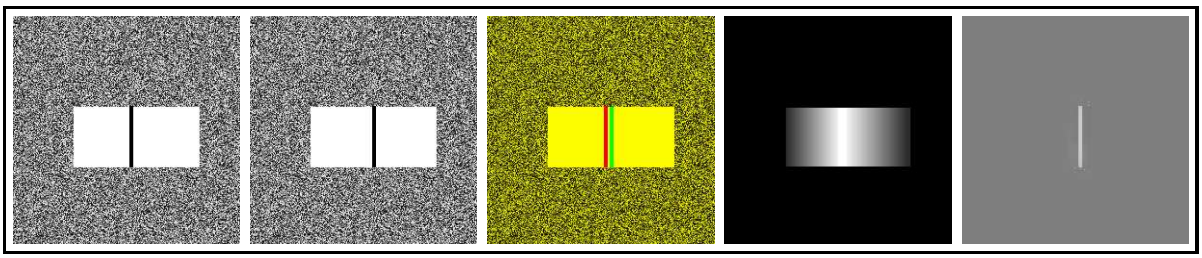


Figure 3.17: *(Left to right) Left image, right image, anaglyph, perceived disparity, estimated disparity map using SOFM.*

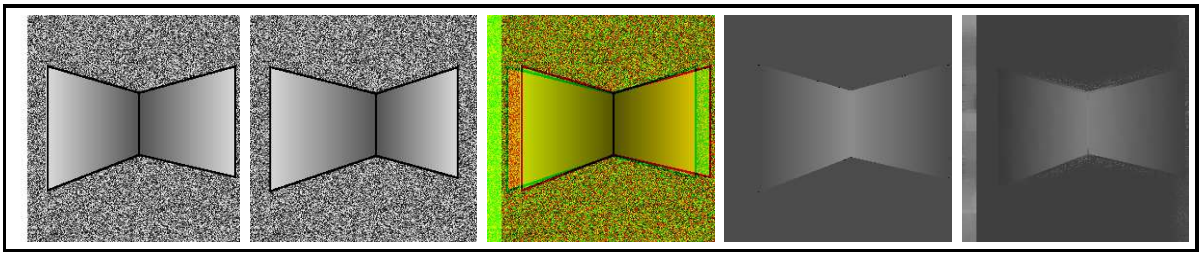


Figure 3.18: *(Left to right) Left image, right image, anaglyph, perceived disparity, estimated disparity map using SOFM.*

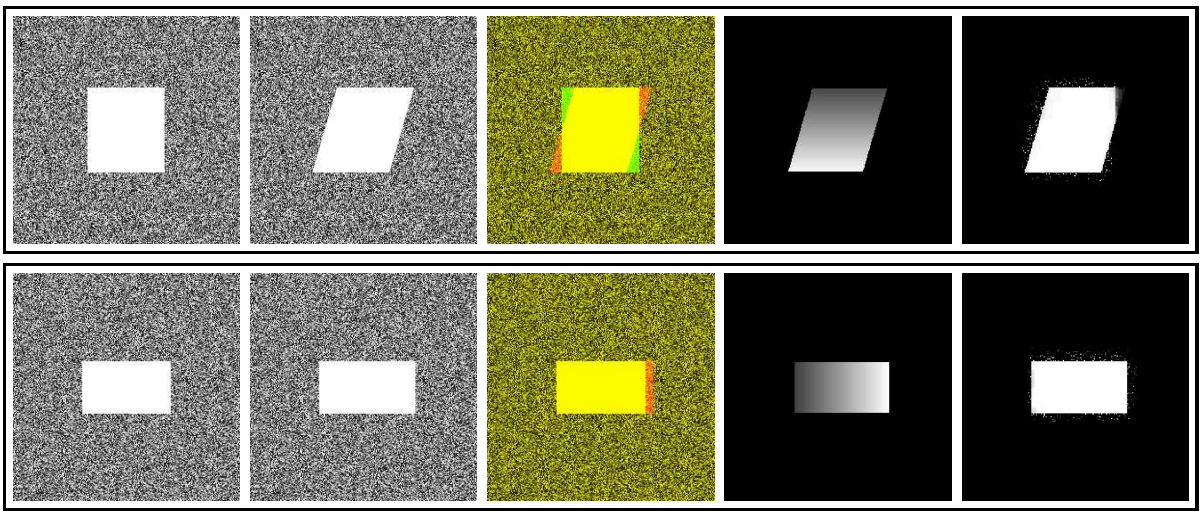


Figure 3.19: *(Left to right) Left image, right image, anaglyph, true-disparity map, estimated disparity map using MSOFM (1st row) Vertical gradient, (2nd row) horizontal gradient.*

3.3.4 Other Stereo-Pairs

(1) **Non-epipolar:** In Fig. 3.20 the right image of the tsukuba stereo-pair is altered to create non-epipolar stereo-pairs. The results in Fig. 3.20 show the superior performance of our approach compared to other algorithms. Hence the MSOFM-based approach doesn't need epipolar constraint. The PBD for the MSOFM (Table 3.4) shows that it is robust with non-epipolar constraint.

Figure 3.20: *(Left to right) Original right image, altered right image and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, MSOFM for (1st row) right image with a vertical shift of 5 pixels, (2nd row) right image 90 % of the original height.*

(2) **Noisy:** The robustness of the algorithms against noise in the stereo-pair is analyzed by estimating disparity map with added noise in the right image. Noise such as blur may be due to defocus of one camera, or CCD thermal noise (impulse noise) or difference in gain of the two cameras. The resulting disparity-maps with impulse noise, and blurred right image are given in Fig. 3.21.

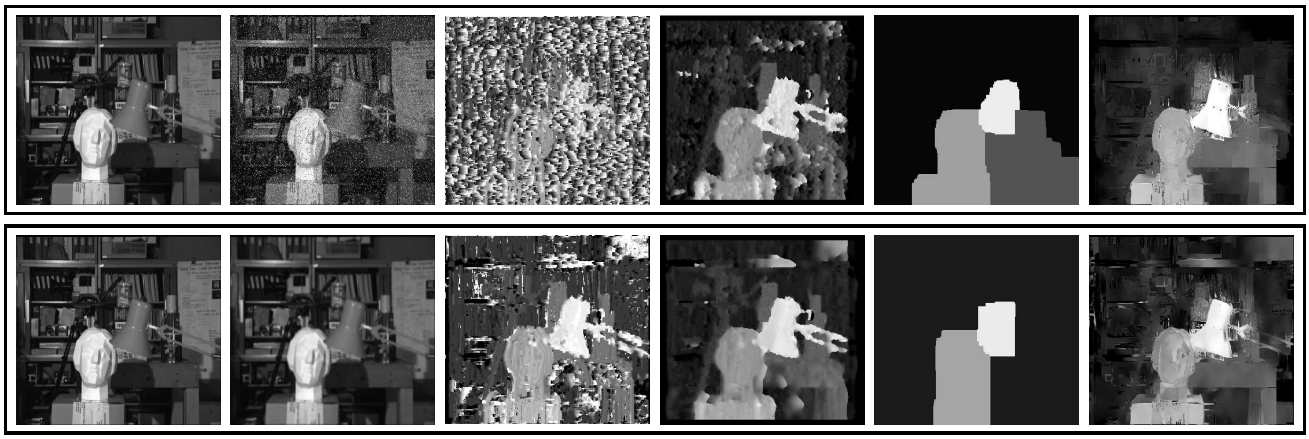


Figure 3.21: *(Left to right) Original right image, right image with noise and the corresponding disparity maps obtained using SSD, cooperative stereo, Graph-cuts, SOFM for (1st row) right image with* 10% *impulse noise (2nd row) right image with Gaussian* ($\sigma = 5$) *blur.*

Since the pixel-by-pixel intensity is used to find the closest match and for updation, if the two images of the stereo-pair have different contrast, the network fails to give an acceptable result (see Fig. 3.22).

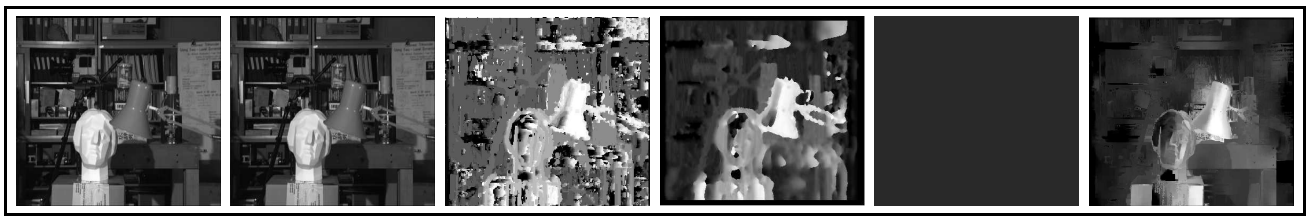


Figure 3.22: *(Left to right) Original right image of the Tsukuba of stereo pair, the contrast enhanced version, disparity estimate of SSD, cooperative stereo, Graph-cuts and MSOFM.*

General Remarks (1) Specular surfaces: Nayar *et al.* [43] present a physical model based solution to stereo with specular reflection. Li *et al.* [44] propose an approach in which the specular regions are isolated before disparity estimation, and then the estimated disparity is used to reconstruct the isolated regions. However, no standard image-pair with specular reflection is available to check the performance of our algorithm. The disparity

map, estimated for the stereo-pair with specular reflection, is shown in Fig. 3.23. Observe that, in the region of reflection, the disparity map is erroneous.

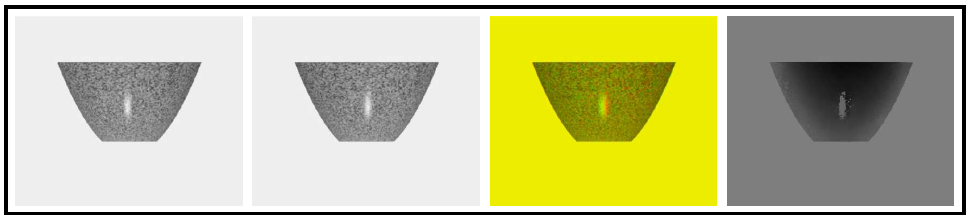


Figure 3.23: *A stereo-pair with specular reflection : Left image, right image, anaglyph, estimated disparity.*

(2) The hybridized model of Belhumeur [8] seems to be an *ad hoc* procedure for dealing with non-local interactions. Compared to other algorithms, the MSOFM has inherent global interaction. It should be interesting to explore the possibility of casting MSOFM in a Markov-field framework.

(3) Our motivation has been the human visual system's handling of occluded regions, salient features in the scene, and global 3-D structures (large background surfaces continuing behind smaller foreground surfaces). But our algorithm does not take into account the geometric structure of the scene. The matching produced is purely based on intensity correlation. Moreover, the distinction between a large background and small foreground is based on heuristic observations.

(4) "The perceived disparity of occluded features is the limit of their possible disparity values (Panum's limit case), if no other source of information is given." [45] In our algorithm, the occlusion is assigned a disparity based on its similarity in intensity to the background or the foreground. But, according to above mentioned conjecture, the occluded region is always assigned the disparity of the background. Hence the conjecture follows the geometry, but our algorithm is purely based on intensity similarity.

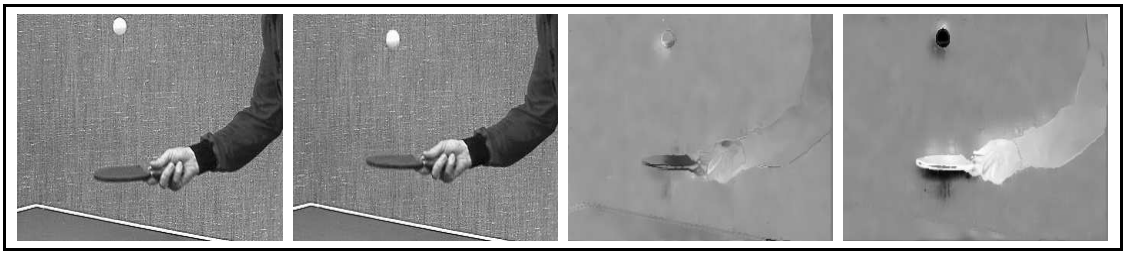


Figure 3.24: *(Left to right) Frame 1 and frame 2 of a video sequence, horizontal and vertical disparity maps.*

3.3.5 Merits of MSOFM

On the basis of the examples given above, we summarize the merits of the proposed approach. (1) As applied to stereo disparity estimation of non-epipolar stereo-pairs, the proposed approach is believed to be novel. (2) There is no need for an interpolation stage to propagate the disparities along horizontal lines, since the MSOFM is able to estimate disparities in these regions by local cooperation. (3) It is possible to extend the approach to color images, by using the YUV components instead of a single intensity. (4) The same approach can be used for motion estimation, as d_x represents the x-component and d_y the y-component of the flow vector (see Fig. 3.24).

3.4 Summary

We have proposed a novel application of the self-organizing neural network (SONN) for estimating the disparity map from a stereo-pair of images. The standard network is modified in order to take into account the properties of stereo matching. After initializing the network to one of the images, we compute the amount of *deformation* required to transform it into the other image. An interesting result is that the deformation is a measure of disparity. The approach has many special features: no assumption is made with respect to epipolar geometry of the images; it can take into account salient features (like discontinuous depth values at boundaries and surface wrinkles) and half-occlusions; and, when applied to both synthetic and natural stereo-pairs (including wire-frames), performs better than existing algorithms. Many examples are given to illustrate its superiority and limitations. Extension of the approach to motion estimation in an image sequence has also been successfully explored.

Chapter 4

Conclusions

The thesis has dealt with the following problem in computational stereo for extraction of three-dimensional scene structure from images:

Given a stereo- (or binocular) image pair of a physical scene, estimate the depth or distance of all the points in the scene as measured from, say, the line joining the centers of the two cameras?

The basis for attempting to find a solution to the above problem is the fact that a single three-dimensional physical point projects to a unique pair of image points in the two cameras. Therefore, given the two camera images (which are slightly different from each other because of the separation of the two cameras), if we can locate the image pixels that correspond to the same physical point, then it is possible to determine its three-dimensional position. The distance between the two image pixels corresponding to the same physical point is called **disparity**. The two points in the two images are said to 'correspond' with each other (and hence are called 'correspondence points').

The amount of disparity is inversely proportional to the depth of that point in the scene. The two points in the two images are said to 'correspond' with each other (and hence are called 'correspondence points').

The present work is motivated by the remarkable characteristics of the human visual system (and of mammals) to perceive depth *effortlessly* from the images formed on their two retinas. Human depth perception is based mainly on the processing of the retinal images by the neurons in the brain, using an architecture which is not yet completely understood. Any model which even crudely imitates the performance of the HVS is believed to provide (i) a non-classical framework to deal with stereo-image analysis and, at the same time, (ii) a possible insight into the HVS itself. This has been the main stimulus to attempt an automatic depth extraction from such images using artificial neural networks.

The stereo-problem is equivalent to the problem of determining the correspondence points from which (i) the disparity at each pixel is obtained directly, and (ii) an image

of disparity values (called the disparity map) can be created. Determination of the correspondence points has turned out to be extremely hard. In an attempt to make the problem tractable, many constraints are assumed in the literature. If the disparity map can somehow be estimated, physical depth from a disparity map can be computed from a knowledge of camera geometry.

The thesis presents some novel applications of (i) cooperative computation and (ii) self-organizing maps to estimating the disparity map from stereo-image pairs of (i) random dot stereograms, (ii) wire-frame models and (ii) natural images.

The main results are:

1. A new model for stereo disparity estimation has been proposed using a single-layer artificial neural network (ANN). The main result is that we determine the network weights using synthetic stereo-image data with known disparity maps. The novelty of the approach lies in the fact that the weights characterize the amount of support from each neighbor of the pixel under consideration.

   These weights suggest a new model of neuronal excitation-inhibition (for stereo-image analysis) that is distinct from other models.

   When it is applied to random-dot and natural stereo image-pairs, the performance of the proposed network compares favorably with that of recent algorithms. However, its superiority consists in (i) imposing no constraint on disparity gradient; and (ii) a non-iterative computation that is fast and that can be implemented in parallel since the compatibility matrix for each pixel can be constructed separately.

2. A modified self-organizing neural network (SONN) is employed to determine the amount of deformation required to transform one of the images to the other. It turns out that the deformation *itself* is a measure of disparity. Some special features of the new approach are: (i) it makes no assumptions with respect to epipolar geometry of the images; and it can take into account salient features (like discontinuous depth values at boundaries and surface wrinkles) and half-occlusions. And when applied to both synthetic and natural stereo-pairs (including wire-frames), performs better than existing algorithms. Many examples are given to illustrate its superiority and limitations.

   Extension of the approach to motion estimation in an image sequence has also been successfully explored.

The cooperative model approach compares favourably with the existing algorithms, but the SONN-based approach is found to be more robust with respect to noise, blurring

and quantization than the algorithms of the literature. Further, it is believed that the results obtained from the SONN constitute the *first demonstration* that the disparity map can be easily and efficiently computed with the help of the self-organization principle. The advantages and disadvantages of the methods are also discussed.

It is now the practice to use multi-camera images and/or video sequences of stereo-pairs in order to create an enhanced feeling of 3D perception. Extension of the two proposed approaches to deal with such datasets in order to arrive at real-time computation of depth-maps is believed to be of significant value in practice, especially in mobile robotic vision. Also a generalization of the cooperative network with supervised learning to deal with occlusion seems to be challenging in itself.

Papers related to the contents of the thesis

1. On the Application of a Self-Organizing Neural Network to Stereo-Image-Pair Analysis, accepted for presentation at the First International Conference on Pattern Recognition and Machine Intelligence (PReMI'05), ISI, Kolkata.

2. Stereo-Disparity Estimation using a Supervised Neural Network, accepted for presentation at the IEEE International Conference Machine Learning for Signal Processing(MLSP2004), Sau Paula, Brazil.

3. Stereo-Disparity Estimation using a Modified Self-Organizing Neural Network, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, May 2005.

4. Stereo-Disparity Estimation using a Novel Combination of Cooperative Model and Supervised Learning, submited to IEEE Transactions on Image Processing, July 2005.

References

[1] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, October 1976.

[2] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of Royal Society London B*, 204:301–328, 1979.

[3] D. Marr. Vision. *Freeman, USA*, 1982.

[4] C. L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:675–684, July 2000.

[5] K. Prazdny. Detection of binocular disparities. *Biological Cybernetics*, 52:93–99, 1985.

[6] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Pmf: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985.

[7] W. E. L. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):17–34, January 1985.

[8] P. N. Belhumeur. A computational theory for binocular stereopsis. Book: *Perception as Bayesian Inference, Edited by D. C. Knill and W. Richards, Cambrdige University Press*, pages 323–364, 1996.

[9] D. G. Jones and J. Malik. A computational framework for determining correspondence from as set of linear spatial filters. *Computer Vision — ECCV'92*, pages 395–410, 1992.

[10] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *IEEE Trans. PAMI*, 25(8):993–1008, August 2003.

[11] T. Shipley. Report on the first random dot stereogram by aschenbrenner in 1954. *Vision Research*, 11:1491–1492, 1971.

[12] B. Julesz. Stereoscopic vision. *Vision research*, 26:1601–1612, 1986.

[13] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, 1:195–202, June 2003.

[14] U. R. Dhond and J. K. Aggarwal. Structure from stereo — a review. *IEEE Trans. Systems, Man, and Cybernetics*, 19:1489–1510, 1989.

[15] J.J. Little and W. E. Gillett. Direct evidence for occlusion in stereo and motion. *Image and Vision Computing,*, 4(8):328–340, 1990.

[16] D. Tzovaras, M. G. Strintzis, and H. Sahinoglou. Evaluation of multiresolution techniques for motion and disparity estimation. *Image Communication*, 6(1), June 1994.

[17] F. Porikli and Y. Wang. Disparity estimation by patch matching. *Proc. Picture Coding Symposium (PCS97), Berlin, Germany*, pages 433–438, September 1997.

[18] T. Kanade and M. Okutami. A stereo matching algorithm with adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–933, September 1994.

[19] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. *Proc. Third European Conf. Computer Vision*, pages 150–158, 1994.

[20] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. *Proc. Int'l Conf. Computer Vision*, 22:492–499, 2000.

[21] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, Nov. 2001.

[22] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14:553–572, 1982.

[23] S.S. Intille and A.F. Bobick. Incorporating intensity edges in the recovery of occlusion regions,. *Proc. Int l Conf. Pattern Recognition,*, 1:674–677, 1994.

[24] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):139–154, March 1985.

[25] H. Maitre and Y. Wu. Dynamic programming algorithm for elastic registration of distorted pictures based on autoregressive model. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(2):288–297, February 1989.

[26] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts,. *Proc. Int l Conf. Computer Vision,*, 2001.

[27] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Microsoft Research Technical Report MSR-TR-2001-81*, November 2001.

[28] M. H. Lin and C. Tomasi. Surfaces with occlusions from layered stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1073–1078, August 2004.

[29] S. D. Cochran and G. Medioni. 3-d surface description from binocular stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 981–994, 1992.

[30] J. M. Hakkarainen, J. J. Little, Jr. H. Lee, and J. L. Wyatt. Interaction of algorithm and implementation for analog vlsi stereo vision. *SPIE Visual Information Processing: From Neurons to Chips*, pages 173–184, 1991.

[31] A. Khotanzad, A. Bokil, and Y. W. Lee. Stereopsis by constraint learning feedforward neural networks. *IEEE Transactions on Neural Networks*, 4(2):332–342, March 1993.

[32] A. Khotanzad and A. Bokil. A constraint learning feedback dynamic model for stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(11):1095–1100, November 1995.

[33] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin. Phase-based disparity measurement. *CVGIP:Image Understanding*, 53(2):198–210, 1991.

[34] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman. Stereoscopic depth discrimination in the visual cortex — neurons ideally suited as disparity detectors. *Science*, 247(4972):1037–1041, 1991.

[35] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *Computer Vision — ECCV'92*, pages 424–433, 1992.

[36] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts,. *Proc. Int l Conf. Computer Vision,*, 2001.

[37] M. Gong and Y. H. Yang. Genetic-based stereo algorithm and disparity map evaluation. *International Journal of Computer Vision*, 47(1/2/3):63–77, 2002.

[38] M. G. Strintzis and S. Malassiotis. Object-based coding of stereoscopic and 3d image sequences. *IEEE Signal Processing Magazine*, 16(3):14–28, May 1999.

[39] Y. V. Venkatesh, K. R. Ramakrsihnan, S. Kumar Raja, and A. Jaya Kumar. Cooperative computation-based neural network for estimating disparity in stereo-image pairs. Technical report, Electrical Department, Indian Institute of Science, Bangalore, India, July 2005.

[40] T. Kohonen. Self-organization and associative memory. *Springer, Berlin*, 1989.

[41] B. L. Anderson and K. Nakayama. Towards a general theory of stereopsis: Binoclular matching, occluding contours, and fusion. *Psychological Review*, 101:414–445, 1994.

[42] K. Nakayama and S. Shimojo. Davinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Research*, 30(11):1811–1825, 1990.

[43] Dinkar. N. Bhat and Shree. K. Nayar. Stereo in the presence of specular reflection. *Int. Conference on Computer Vision,*, 1995.

[44] Y. Li, S. Lin, H. Lu, Sing Bing Kang, and H.-Y. Shum. Multibaseline stereo in the presence of specular reflections. *ICPR*, 2002.

[45] D. Weinshall and J. Malik. Review of computational models of stereopsis. *Book: Early Vision and Beyond, T. V. Papathomas (editor), MIT Press*, pages 33–41, 1995.