National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

FAST IMAGE SEGMENTATION USING STEREO VISION

Majid Ezzati

Department of Electrical Engineering
McGill University

September 1995

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements of the degree of
Master of Engineering

ISBN 0-612-07974-0

Abstract

Binocular stereopsis is a biologically motivated approach that uses two slightly different views of a scene to extract information about its three-dimensional properties. The two underlying principles of our approach to stereo vision are local computation of binocular disparities and the use of the resulting disparity map for image segmentation.

The cepstrum is used to provide an estimation of binocular disparity between corresponding regions of the stereo image pair. We study the cepstrum and its properties, and suggest improvements to the initial disparity estimation stage. Next a modified median filtering scheme is employed for the refinement of the initial disparities using neighbourhood information. The overall disparity map is used for image segmentation based on distance.

Local estimation of initial disparities provides two fundamental advantages for real-time systems: the possibility of increased computational efficiency through parallel implementation and a fixed running time that is independent of image properties. Furthermore, using stereopsis for figure-ground segmentation rather than surface reconstruction eliminates the need for camera calibration which is essential for methods based on exact depth calculation. Therefore, the approach is well-suited to active vision systems in which the cameras are in constant motion.

We provide evidence for the plausibility of the disparity estimation algorithm and the properties of the overall disparity map in the context of biological stereopsis. The algorithm is implemented on a network of TMS320C40 processors to obtain a processing time of one second for a 128-pixel $\times$ 128-pixel image frame.

Résumé

La stéréovision binoculaire est une approche biologique qui utilise deux vues légèrement différentes d'une scène pour en extraire des propriétés tridimensionnelles. Les deux principes sous-jaçents de notre approche de la stéréovision sont le calcul local des disparités binoculaires et l'utilisation de la carte de disparité résultante pour segmenter l'image.

La technique du "cepstrum" est utilisée pour obtenir une estimation de la disparité binoculaire entre les régions correspondantes des deux images. Nous avons étudié le "cepstrum" et ses propriétés, et suggérons des améliorations au processus d'estimation de la disparité initiale. Par la suite, nous utilisons un filtre médian modifié pour affiner l'estimation initiale de la disparité grâce à une information locale. Le résultat de la carte de disparité est alors utilisé pour réaliser une segmentation d'ímage basée sur la distance.

L'estimation locale des disparités initiales offre deux avantages majeurs pour une estimation temps réel: la possibilité d'accroître l'efficacité calculatoire grâce à une implémentation parallèle, et un calcul à temps constant indépendant de la complexité de l'image. Un autre avantage est que l'utilisation de la stéréovision pour la segmentation "avant-arrière plan" plutôt que pour la reconstruction de surfaces élimine le besoin d'une calibration de caméra. Ainsi cette approche est adéquate aux systèmes de vision active pour lesquels la caméra est en mouvement constant.

Nous montrons ainsi dans un contexte biologique, la plausibilité de notre algorithme d'estimation de disparité et de ses propriétés pour l' obtention d'une carte globale de disparité. L'algorithme a été implémenté sur un réseaux de processeurs TMS320C40 et permet d'obtenir une image par seconde pour une taille d'image de 128×128 pixels.

Acknowledgements

This thesis will not be complete without thanking the many people whose presence, help, and support I have cherished over the past two years.

First and foremost I shall thank my advisor, Professor Martin Levine. His encouragement to challenge new frontiers, accompanied by his constant help, caused the experience of the the past two years to be beyond a mere academic endeavour. I am especially grateful to him for his patience and support during the months when making a decision about the future seemed confusing. I also wish to thank Ms. Ornella Cavaliere for her encouraging kindness, especially during the stressful moments.

Some of my most enjoyable moments in Montreal and at McGill were those shared with friends and colleagues. I am thankful to them all. I specifically thank Marc Bolduc and Gal Sela for their enormous guidance, help, and patience which made the implementation stage of this work possible. If Marc ever teaches a course in philosophy of computing, I will be the first to attend. I wish to thank Paul Mackenzie for the careful reading of the initial draft of this thesis and valuable feed-back. I also thank Dr. Thierry Baron for the French translation of the abstract. Salutes to (soon-to-be-Drs) Farzam Ranjbaran, James Elder, Kenong Wu, and Michael Kelly for the valuable advice and the sharing of their experiences and wisdom when things looked unclear.

I shall also thank those who, although not directly involved in my graduate studies at McGill, made it possible to take on the experience. I am thankful to Professor Max Wong of McMaster University. Working with Professor Wong introduced me to the joys of independent work and taught me valuable lessons about it. I appreciate his help and support which continued well after I had left McMaster. Many thanks to Adriane Weller for her help from the time that I considered applying to McGill to the last days of work on this thesis. Adriane's presence and friendship made the move to McGill much easier and living in Montreal much more pleasant. Most importantly, I wish to thank my family for their ever-present support and kindness. My brother, Saied, has become the person I

have learned to depend on the most. Without him, I simply would not have made it! My parents' dedication to my education and success has been the motivation that could get me through the toughest moments. I am grateful to them for ever. To them, is this thesis dedicated.

I would like to express my gratitude to the Natural Sciences and Engineering Research Council (NSERC) of Canada for financial support through a Post-Graduate Scholarship. This research has also been supported by funding through the Institute for Robotics and Intelligent Systems (IRIS) as a part of the National Centres of Excellence Program.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

By projecting a three-dimensional scene onto a two-dimensional image plane, an imaging system forfeits all *direct* knowledge about distance or depth. Any information about the third dimension is embedded *indirectly* in the two-dimensional image(s) of the scene. Depth cues are represented as relationships either among the intensities of a single image or between the intensities of multiple images of the scene.

Due to this lack of direct information about distance, the determination of three-dimensional structure has been among the most challenging problems in computer vision. The wish to obtain knowledge of the third dimension has on one hand led to active range-finding techniques. On the other hand, it has initiated efforts to reconstruct depth from the knowledge embedded in the two-dimensional images. Active range-finding uses the reflection of a known moving light source projected on different points of the scene, along with triangulation, to determine the distance of these points [**12**]. An example of such a technique uses laser technology.

Biological vision systems seem to have an outstanding grasp of the three-dimensional world using the two-dimensional images projected onto the retinae. This has further strengthened the efforts to obtain distance information from two-dimensional images. This research has produced a class of surface extraction schemes known as "*shape from X*" methods. These algorithms attempt to exploit the relationships among the intensities of one or more images caused by the three-dimensional nature of a scene. In other words, the methods attempt to reverse the process causing such relationships in order to obtain the underlying three-dimensional structure. Shape from shading algorithms such as those described in [**49**], [**90**], and [**22**] are examples of methods which use the intensities of a single image to recover the third dimension. Shape from motion and shape from stereo use multiple images

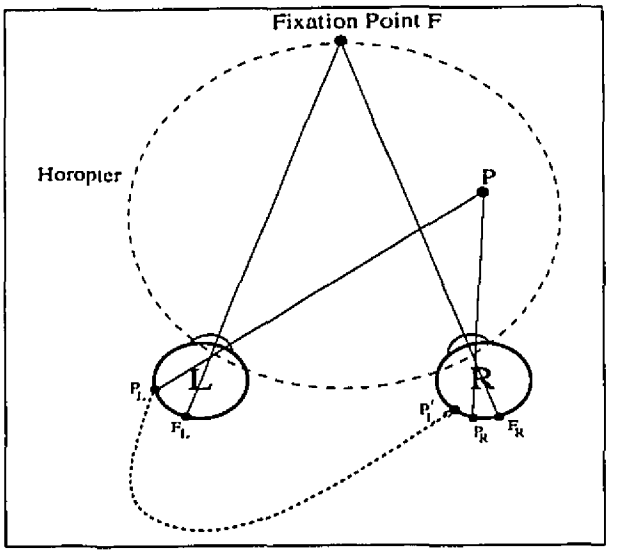


FIGURE 1.1. Binocular Disparity. F_L and F_R represent the images of the fixation point, F, on the left and right eyes respectively. P_L and P_R represent the images of point P on the left and right eyes, respectively. P'_L indicates the point with the same coordinates as P_L on the right retina. The difference between P_R and P'_L is the disparity of point P.

of the scene to infer its three-dimensional structure. In particular, stereo vision is often also motivated by the binocular visual system of primates.

1. Stereo Vision and Binocular Disparity

The point of intersection of the axes of two imaging devices is known as their point of fixation. Geometrically, all points lying on a locus passing through the fixation point are projected onto identical locations on the left and right image planes. The image of any point nearer or farther than this locus, referred to as the horopter, is formed at different locations on the two retinae. The difference between the projections of a point on the left and right retinae is known as the binocular *disparity* of the point. Figure 1.1 is an illustration of the concept of binocular disparity.

For any given camera set-up, the disparity of a point is dependent on its distance from the horopter. Therefore, the locus of a point in depth can be obtained from its disparity and a knowledge of the camera arrangement. Given such a relationship between disparity and depth, surface reconstruction using stereo vision can be divided into two separate problems: first, obtaining the vector displacement or disparity of corresponding sections of the two images; and second, using the disparity information to obtain the three-dimensional

knowledge of the scene. Such an approach has formed the basis for most of the traditional schools of thought on stereopsis.

Obtaining the precise disparity profile of a stereo image pair has resulted in algorithms that not only are computationally expensive but also have image dependent running times. The second stage of surface reconstruction using stereopsis poses another set of problems. Inferring absolute distance from disparities requires accurate and detailed knowledge of the camera set-up. Grimson shows that small deviations from such knowledge can result in considerable errors in the calculation of the absolute depth [**42**]. Despite the fact that they can produce relatively accurate distance profiles under highly structured conditions, the traditional stereo surface reconstruction methods are of little use for complex robotic vision systems acting in dynamic environments.

2. Motivation for an Alternative Approach

Many applications of visual perception involve dynamic scenes whose attributes change constantly. Such a change may be due to the motion of the objects in the scene, the observer, or both. In recent years computer vision has observed the emergence of active vision as a response to the problem of enabling autonomous agents to deal with the changing nature of their environment [**3**], [**5**]. An active vision system is one whose physical configuration varies in response to the changes in the scene or the required visual information. Although the criteria which guide the particular control mechanism may differ from one system to another, all active vision systems share the property of being influenced by and "adapting" to the evolution of their environment. With such a characteristic, active vision systems require algorithms whose computational requirements reflect the rate at which the visual information changes. Furthermore, such computational requirements should be independent of specific image properties.

Another influence on both the fields of biology and computer vision has been the observation that visual perception is the *overall* result of multiple visual cues as well as their interactions [**23**], [**28**], [**115**], [**117**], [**118**]. This is in contrast to the view that each low level feature is required to result in a complete description of the scene, independently from other such cues. In a system which uses multiple visual features, the perception of the scene resulting from each individual visual cue may be incomplete. Instead, the visual system probably obtains the best possible description of the scene using all available visual signals in a manner which is appropriate for the functioning of the agent. Figure 1.2 is a schematic illustration of such a multi-feature visual system coupled with the concept of

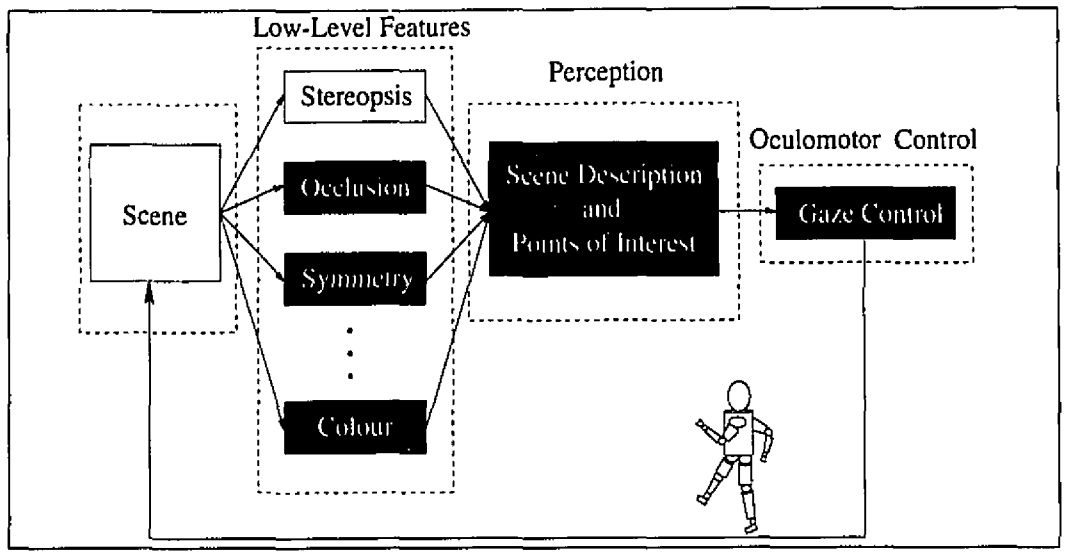


FIGURE 1.2. Multi-Feature Foveated Active Vision System for a Dynamic Environment.

active vision. In this thesis, we consider stereopsis as one of the multiple features of such an active vision system. In doing so, we attempt to adhere to the principle that the overall perception is a result of both the individual components and their inter-relationships.

In studying vision, first and foremost one must remember that there is a purpose for every visual perception system. The characteristics of the system should reflect and be in harmony with its intent. The objective of the system considered in this work is to guide an autonomous mobile robot in an unstructured environment and enable the robot to perform specific tasks. Functioning in such an environment demands the ability to perform certain operations such as obstacle avoidance and object recognition. It also requires that the actions taken by the robot are decided upon and executed within a reasonable period of time.

The visual system considered for this purpose uses active vision. This enables the agent to adapt to the changes in its environment. It also allows for filtering and reducing the received visual information [**19**] to a subset which is of interest to the system [**110**]. This visual system uses multiple visual cues to obtain a complete description of its environment. Stereopsis and stereo disparity form one such cue, used to provide information about the relative depths in the scene. Distinguishing between surfaces that are located at different distances provides information about the available paths. Similarly, distinguishing an object from its background forms the first stage of object recognition in a complex scene. Therefore,

figure-ground separation on the basis of (difference in) distance is of value to the visual system of a mobile robot. It is also important to ensure that the time required for performing the operation is short and independent of specific scene properties.

3. Our Approach to Stereopsis

Reduced response time can be achieved by parallel processing of visual information. One can further reduce the processing time by assigning a single value as the output of a specific operation, such as stereo disparity estimation, to small regions of the image rather than individual image points. We provide evidence that the above strategies in fact exhibit a plausible level of similarity to those of the early stages of stereo disparity estimation in biological systems.

This thesis takes an approach which is unlike that of many traditional stereo algorithms. These algorithms, motivated by the wish to construct a precise depth map, involve a sequential search to find a match for every image feature or point. The dislocation of each feature determines its disparity. To ensure the global consistency of the matched elements, most such algorithms also involve an optimization stage at every iteration.

We consider initial disparity estimation as a local computation which can be implemented in parallel. This produces increased processing speed as well as a running time which is independent of image properties. The disparity estimation stage is based on the algorithm originally presented by Yeshurun and Schwartz [**125**]. The method divides the two images of the stereo pair into small sections and obtains initial estimations for the disparities of all such sections. The division of the image pair is motivated by data representation in the ocular dominance columns of the primary visual cortex. There information from the left and right eyes are represented in the form of interlacing image "patches". The algorithm uses the cepstrum, a method traditionally employed in echo detection, and provides a local estimation of binocular disparity between corresponding patches.

We study the properties of the cepstrum and use them to suggest improvements to the disparity estimation algorithm of [**125**]. Furthermore, by carefully examining the performance of the cepstrum on various types of signals, we infer that the algorithm may result in the detection of false disparities at some ocular dominance columns. To deal with this issue we have developed a method for refining the raw or initial disparity map obtained by estimating disparities using the (improved) cepstral filtering algorithm. Next we refine the initial disparity estimates using neighbouring disparity information. We employ a modified median filtering scheme for this purpose.

The final depth map contains information about the three-dimensional properties of the surfaces in the scene. However, its most salient features are the discontinuities in distance which mark the boundaries between various surfaces and objects in the scene. We assume that the refined disparity map is used for differentiating between surfaces that are located at different distances and not for precise surface reconstruction. Using stereopsis for figure-ground segmentation, rather than surface reconstruction, also eliminates the need for camera calibration which is essential for exact depth calculations. Therefore, the approach is well-suited to active vision systems in which the cameras are in constant motion. Psychophysical evidence is provided in support of the fact that such a symbolic representation is in agreement with the properties of human stereopsis.

4. Contributions

The following are the main contributions of this thesis:

- Analyzing the influences of the dimensions of the ocular dominance columns on the algorithm. This is used for choosing the dimensions of the columns which are taken from the stereo image pair.
- Examining the properties of the cepstrum - the disparity estimation tool - and formalizing the dependence of its performance on the signal properties. This permits us to suggest modifications that compensate for some of its shortcomings.
- Relating the concepts of stereo disparity, the local estimation strategy, and the properties of the cepstrum to the occurrence of false disparities.
- Presenting a method for the refinement of the initial disparity map to produce a scene representation appropriate for depth-based image segmentation.
- Relating the technique and the final disparity map to the neurophysiological and psychophysical properties of biological vision systems.
- Parallel implementation of the method to obtain a processing time of one second for a 128-pixel $\times$ 128-pixel image.

5. Overview and Organization

The next two chapters briefly review computational stereo vision and the significant findings in biological stereopsis. Chapter 4 along with Appendices A, B, and C study the cepstrum and its role as a tool for echo detection. Then Chapters 5 and 6 explain the initial disparity estimation and disparity map refinement stages of the overall method respectively.

Experimental results are provided in Chapter 7 followed by evidence for biological plausibility of the approach in Chapter 8. Next the details of parallel implementation are presented. Finally, Chapter 10 serves to state the concluding remarks and directions for future work.

CHAPTER 2

Computational Stereo

From the definition of binocular disparity, one can deduce that for a given camera set-up, the knowledge of the disparity of a point in the scene determines its locus in depth. Combining the depth locus with the coordinates of the point in one image of the stereo pair provides the precise location of the point in space. The set of all the disparities of an image pair is known as its *disparity map.* Notwithstanding occlusion, it is therefore possible to obtain the complete three-dimensional structure of a scene without ambiguity from the knowledge of camera set-up and its complete disparity map - a long-awaited goal in computer vision. Computing three-dimensional structure from the disparity map and imaging geometry is a well-defined geometric reconstruction problem. Nevertheless, obtaining the complete and correct disparity map of the scene is a challenge when attempting to use stereopsis for depth reconstruction.

1. The Correspondence Problem

A possible approach to obtaining the disparity map of a stereo image pair is to choose a set of features or elements in one image, and determine the corresponding elements in the second. The displacement between the respective positions of a feature in the two images equals the disparity of the feature. The problem of finding matching elements in one image for elements of the other image is known as the *correspondence problem.* In determining correspondence, stereo algorithms make assumptions about the surfaces in the world and the imaging process. These assumptions are described by Marr in [**71**] and [**98**] as follows:

- Compatibility assumption: All objects in the scene result in similar features in the left and right images. Therefore the two images of a stereo pair should embody a certain level of similarity. This similarity forms the underlying principle of the

matching process. If two pixels, regions, or edges have arisen from the same entity in the scene, they will match; if they have not, they cannot be matched.

- Uniqueness assumption: Almost always, a feature in one image can match to no more than one feature in the other. Therefore, only a single disparity can be assigned to each image feature.
- Smoothness assumption: The disparity of the features varies smoothly almost everywhere in the image. This is equivalent to assuming piecewise continuous surfaces in the scene.

In addition to the above three assumptions, stereo algorithms sometimes employ one or more of the following constraints in solving the correspondence problem:

- Viewing geometry constraint: Corresponding features lie on a locus determined by the geometry of the imaging devices. This locus is referred to as the *epipolar line* [**56**]. This assumption simplifies the search for corresponding features by limiting it to epipolar lines.
- General position constraint: Events that occur quite infrequently, in a statistical sense, do not result in corresponding elements [**67**]. For example, corresponding image features are required to have the same ordering or arrangement in both images of the stereo pair [**6**].
- Disparity gradient constraint: The rate of change of disparity across the image is limited [**74**]. This constraint, of course, further enforces the smoothness assumption.
- Maximum detectable disparity : The disparities of a stereo pair lie within a limited range, or equivalently, the range of distances detectable by the system is limited.

The role of the above assumptions is to simplify the search for the correct solution to the correspondence problem by involving constraints about the scene and imaging process in the search.

2. Solving the Correspondence Problem

The simplest set of image features, for which the correspondence problem is defined, is the set of individual image samples or pixels. Using pixels for determining correspondence is equivalent to obtaining the disparity map by determining the projections of every point in the scene into both images and measuring the displacement between their respective positions. However, images are generally represented using a finite set of grey levels whose membership is much smaller than the total number of image samples. Consequently,

determining correspondence by merely matching single pixels is inherently ambiguous. To avoid such ambiguity, stereo algorithms obtain the disparity map using methods other than direct matching of individual pixel grey levels. Traditional stereo algorithms take two different approaches to the correspondence problem resulting in two distinct classes of stereo algorithms.

2.1. Region-Based Stereo Algorithms. The first class of algorithms, known as region-based algorithms, solves the correspondence problem for individual pixels. However, it uses the grey levels of all the pixels around the target and candidate, rather than just their own, as the measure of similarity between the two points. In other words, these algorithms compare the region surrounding a (target) pixel in one image with the similar area surrounding each candidate pixel in the other. The set of the candidate pixels is likely to be determined based on one or more of the constraints mentioned above. The neighbourhood that most resembles that of the target belongs to the correct match.

Region-based algorithms define some statistical measure on the neighbourhood whose value determines the similarity of the two neighbourhoods and whose minimum (or maximum) indicates the correspondence of the two pixels. For example, such a measure is defined as the sum of squares of differences (SSD) between the corresponding gray shades in windows around the two pixels in the left and right images in [**84**] and [**101**]. Similarly, the normalized mean-squared differences of the grey level values in the two windows is used in [**35**]. Correlation based algorithms, such as those in [**36**], [**37**], [**43**], [**44**], [**45**], and [**65**], use the correlation coefficient between the regions surrounding the candidate points in the image pair to determine correspondence. Moravec [**79**] also uses a similar measure for similarity but performs the matching only among a set of *feature points* in each image. The feature points are selected using an interest operator which, in each neighbourhood of the image, chooses the point whose neighbouring pixels demonstrate relatively high variance. High variance ensures that the correlation coefficient is an effective tool in measuring similarity.

2.2. Feature-Based Stereo Algorithms. The second group of stereo algorithms, known as feature-based algorithms, attempts to solve the correspondence problem for image features more complex than an individual pixel. Increasing the complexity of the image feature reduces the ambiguity of the matching process. Feature-based algorithms implicitly use the similarity constraint. They assume that a complex image feature represents a scene property which produces similar attributes in both images, even though individual image

intensities may be different. Increasing the complexity of the feature increases the likelihood of its uniqueness and perhaps invariance under different viewing conditions.

Edge segments are chosen as image features that are matched in [**30**], [**80**], and [**4**]. Similarly Marr and Poggio [**69**], [**70**], [**71**], Grimson [**40**], [**41**], and Mayhew and Frisby [**74**] choose the zero crossings of the Laplacian of the Gaussian. Nishihara replaces the zero crossings of the Laplacian by its sign which reduces sensitivity to noise [**82**]. "*Region-Based Stereo Analysis for Robotic Application*" [**67**], despite its title, also offers a feature based algorithm in which the *features to be compared* are regions of the image that result from the segmentation of the two monocular images. Barnard and Thompson [**10**] use the output of a modified version of the Moravec interest operator [**79**] as the image feature. Finally, some algorithms use multiple features to characterize image points with rich and highly specific markings and reduce the matching ambiguity. Examples of such approaches are the algorithms in [**59**] and [**56**] which use the response to a set of spatial filters at various orientations, phases, and scales to provide a vector of features for matching.

2.3. Matching Using Features Versus Matching Using Regions. In common between both groups of algorithms described above is a sequential search to find the best match for each image point or feature. The basis for matching in region-based algorithms are the grey levels of regions in the two images, and in feature-based algorithms the extracted features from each image. In other words, region-based stereo algorithms determine correspondence from image intensities and use a numeric representation of the image. On the other hand, by increasing the complexity of the matching kernels, feature-based algorithms solve the correspondence problem by employing a more symbolic representation of the image.

An important problem when using feature-based algorithms is choosing an appropriate image feature which can provide a balance between uniqueness, reliability, density, and computational efficiency. Since in general the density of any image feature more complex than individual pixels is less than the resolution of the image, feature-based algorithms result in sparse disparity maps. However, an advantage of the reduction in the number of matching kernels is higher processing speed of feature-based algorithms *after* feature extraction is completed.

Region-based algorithms, on the other hand, are in essence capable of finding a match for every image point and therefore produce a denser disparity map. However, matching regions for surfaces that have limited texture or in the presence of occlusion can be unreliable. Furthermore, image grey levels which are the basis of the matching process in region-based

algorithms are more sensitive to illumination conditions and viewing angle than complex image features. Therefore, the same physical surface may be represented by different grey levels in the two images due to perspective projection or differences in illumination caused by different viewing directions.

2.4. Imposing the Physical Constraints. To solve the correspondence problem more reliably and efficiently, most stereo algorithms attempt to improve upon basic feature-based or region-based matching schemes. Levine *et al.* deal with the problem of lack of texture by choosing the size of the correlation window proportional to the inverse of the variance of the image grey level in the region [**65**]. Other algorithms use a multiple scales to reduce sensitivity to noise. For example, Grimson [**40**] and Marr [**70**] use a coarse-to-fine approach in which the disparities obtained at larger scales guide the search for the matching feature at the finer stages. Mayhew and Frisby [**74**] use cross-channel correspondence and require that various spatial frequency channels support the disparity of a feature within a certain range.

The most common approach to increasing the reliability and performance of matching algorithms is exploiting one or more of the matching constraints to refine the results of matching. Almost all stereo algorithms use the viewing geometry and maximum detectable disparity constraints to limit the search to a specific range on the epipolar lines. Also a considerable majority of stereo algorithms use one or more of the remaining constraints, particularly compatibility, uniqueness, and smoothness assumptions to reject false matches. Many such algorithms use a measure of the overall quality of matches and perform an iterative and sequential search. In each iteration, the disparities are updated to optimize the measure of quality.

Cooperative stereo algorithms use an approach analogous to the relaxation labeling process of [**52**] or [**100**] to allow the possibility of multiple matches for a single feature at any point during the optimization process. Each possible match is also assigned a likelihood which is updated to reflect the compatibility of the disparity with those at the neighbouring points. At the end of the optimization process, the match with the greatest likelihood indicates the disparity of the feature. Examples of such algorithms are those in [**10**], [**61**], and [**68**]. An important difference between these implementations and classical relaxation labelling is that in [**52**] and [**100**], unlike the above algorithms, the process which provides the initial estimates of likelihoods is distinct from that used to refine these values. A similar approach is taken in [**9**] which uses optimization using simulated annealing to impose the matching constraints. Regularization is used for imposing the constraints in [**96**]. A

dynamic programming approach to solving the correspondence problem is taken in [6]. It uses the Viterbi algorithm to impose a particular illustration of the general position constraint, with the assumption that a left-to-right ordering of edges is preserved along epipolar lines.

Some stereo algorithms use other considerations besides the traditional physical assumptions to improve the matching process. The methods in [35] and [56] require the disparities, and hence the matches, obtained for one of the images of a stereo pair to be consistent with those obtained for the other image. This consistency requirement reduces the possibility of false matches in situations such as occlusion, where no matches for some image points exist. Also clues such as vergence, focus, aperture, and calibration are used to improve on surface estimation using disparity in [1].

A region-based algorithm which uses the sum of squares of differences forms the first stage of a maximum likelihood (ML) algorithm in [73]. Finally, a maximum likelihood approach to solving the correspondence problem in [11]. This paper goes beyond most previous approaches in terms of motivation and states that "the task of a stereo algorithm should be not to simply construct a depth map, but to construct a detailed map of the scene geometry" [11].

Despite many differences, both region-based and feature-based algorithms have one common characteristic. They all attempt to find the match for a feature, be it simple or complex, in one image among a set of candidates in the other by performing a sequential search. This methodology is based on the belief that "matching is a natural way to approach disparity analysis. Assigning disparity classifications to points in a sequence of images is equivalent to finding a matching between sets of points from each image" [10]; or that "an important task for a stereoscopic mechanism is to obtain correct matches between the points in the left and right image, so that the disparity information can be extracted" [46]. Also, most such algorithms impose the matching constraints by optimizing a measure of goodness of the overall matching process. This optimization, in turn, results in iterative algorithms in which the measure of goodness guides the sequential search for the best match. Because of their sequential and iterative nature, most such algorithms are computationally expensive and not suited to real-time applications. Besides the computational cost, another disadvantage of iterative algorithms for real-time applications is the image dependent running time of the optimization process.

3. An Alternative Approach to Disparity Estimation

In recent years, there has been an emergence of a new class of disparity estimation algorithms based on the observation that the spatial shift between the right and left images of a stereo pair results in certain joint spectral and statistical properties. It is such properties, rather than a search, that these algorithms exploit to obtain the disparity map of the image. Although not undermining the possibility of disparity estimation by direct correspondence determination, these algorithms illustrate that matching is not necessary for disparity estimation.

Phase-based disparity estimation exploits the Fourier shift theorem and observes that the phase difference between a signal and its shifted version, at any frequency, is proportional to the spatial shift generating the latter. Phase-based algorithms attempt to measure the difference between the phases of the left and right images. Fourier-based methods simply subtract the left phase and right phase at a given frequency to extract disparity. The signal is multiplied with a rectangularly windowed sinusoid before Fourier transformation in [**122**]. Some phase-based stereo algorithms obtain the disparity from the difference of the phases of the outputs of local bandpass filters applied to the image pair. For example, the algorithms in [**33**], [**34**], [**63**], and [**102**] measure the phase difference of the responses of bandpass Gabor filters to the left and right images. Convolution with Gabor filters is equivalent to replacing the rectangular window of [**122**] by a Gaussian window. The response of the filter reveals information about the phase difference between the original left and right images by providing the variation in the phase relative to the signal at the bandpass frequency of the filter [**53**]. Fleet and Jepson further illustrate that phase information is stable with respect to typical variations between the left and right images such as scale perturbations, smooth shading or lighting variations [**33**], [**34**]. Stability of phase information refers to the fact that small deformations in the image result in only small deformations in the phase signal. Therefore, such an approach is more robust to the differences between the two images of the stereo pair than the traditional region-based matching algorithms. Fleet and Jepson also mention that feature-based matching using the zero crossings of filter responses is analogous to disparity measurements by matching the phase signal at specific image points only [**33**]. These points are, of course, the points where zero crossing information exists. Using such an analogy, they state that phase-based disparity measurement exploits all the phase information, rather than its zero values only, and produces a denser disparity map [**33**], [**34**].

Finally, the fact that the correlation between a signal and its shifted version contains a peak at the location of the shift has given rise to another class of algorithms. These algorithms attempt to obtain the disparity of a stereo image pair by measuring *the location of* the peak in the correlation function between windows from *identical* positions in the two images. The location of the peak of the correlation function, of course, indicates the shift between the two windows. It is important to note the distinction between this class of algorithms and traditional region-based algorithms. Unlike traditional stereo algorithms, correlation-based algorithms do not conduct a search and optimization to find the best match. They measure the disparity directly by determining the *location* of the peak of the *whole* correlation function. Region-based algorithms, on the other hand, perform a search for the match to an image point among a set of candidate points and choose the candidate point whose neighbouring region is most correlated with the neighbouring region of the original point. In other words, the region-based algorithms assume that a number of points may be the match to the original and then eliminate all candidates but one. The elimination criterion is the *magnitude* of the correlation function at a specific point.

Experiments performed in [**55**] illustrate that the performance of correlation for signals with non-white spectra is very poor. Yeshurun and Schwartz [**125**] replace correlation by cepstrum, a technique traditionally used for echo detection, in disparity estimation. Jones and Lamb [**55**] propose some modifications to the traditional cepstrum and use a multiple aperture camera to superimpose the left and right images of the stereo image pair. Multiple evidence is used to increase confidence in the location of the cepstral peak in [**7**].

The emergence of phase-based and correlation-based stereo algorithms has altered the fundamental belief that stereo disparity estimation inherently requires solving the correspondence problem by conducting a search. These algorithms illustrate that the definition of stereo disparity by itself provides other methods for its estimation. It is this basic definition of disparity and the resulting properties of a stereo image pair which this thesis attempts to exploit . Based on this, we propose a sequence of steps for obtaining the disparity map of a stereo image pair.

CHAPTER 3

Biological Stereopsis

The invention of the random-dot stereogram permitted the separation of binocular disparity from other indications of depth such as monocular cues and the presence of known objects in the scene. Random-dot stereograms are image pairs in which one image is a random dot pattern and the other is composed of shifted copies of different regions of the first image. Although there are no monocular depth cues in either image, the disparity resulting from the shifts between corresponding regions of the image pair constitutes a depth cue visible only to binocular viewing of the image pair. The use of stereograms proved Wheatstone's observation that horizontal disparity is sufficient to provide a sensation of depth [**58**]. Since then, it has been shown that other features such as vertical disparity or monocular cues may also be involved in the perception of depth, particularly absolute depth [**75**], [**99**], [**116**]. Furthermore, the use of random-dot stereograms illustrated that image contours are not essential for detecting disparity.

Around the same time as the invention of the random-dot stereogram, Hubel and Wiesel pioneered the study of the "functional architecture" of the primate visual cortex. Their studies, as well as later ones, have illustrated the existence of neurons responsive to binocular disparity in various areas of the visual cortex. Since the discovery of disparity-sensitive neurons many computational theories for the process(es) underlying the responses of such neurons have been offered. Also psychophysical experiments have revealed many of the properties of human stereopsis. This chapter provides a summary of the neurophysiological and psychophysical properties of stereopsis plus an overview of the cortical regions where disparity-sensitive neurons are most observed and projections to and from such regions. One should note that associating disparity-sensitive neurons with certain cortical areas by no means implies a functional specialization for such regions. In addition to disparity sensitivity, the neurons in all of these regions have other properties. In the same manner that

their disparity sensitivity constitutes a response to retinal disparity, their other properties are likely to be responses to other aspects of visual stimuli. Furthermore, there may be other neurons involved in stereopsis whose responses do not resemble those of disparity-sensitive neurons.

1. Ocular Dominance Column Structure and Cortical Projections

Visual data from the Lateral Geniculate Nucleus (LGN) enters the visual cortex at layer 4C of the Visual Area 1 (V1) [**51**]. This data flow is already divided into magnocellular and parvocellular pathways, the former arriving at layer 4Cα and the latter at layer 4Cβ [**50**]. In layer 4Cα, the entrance of the magnocellular pathway to the cortex, the visual data is organized into interlacing columns from the left and right eyes. This form of representation is known as the ocular dominance column structure of the visual cortex. The columns corresponding to each eye provide a topographical map of the retina and the visual field. *In cortical units* all columns have equal width. On the other hand, *in retinal units* or *in the units of visual angle*, the columns correspond to larger areas of the retina with increasing eccentricity. In other words, with increasing distance from the fovea a larger portion of the retina, and equivalently of the visual field, is mapped into the same area of the cortex.

The receptive fields of the neurons of layer 4Cα are all of the circle-surround form found in the earlier stages of the visual pathway [**50**], [**51**], [**64**]. Each ocular dominance column probably receives inputs from the same number of LGN fibres and contains the same number of receptive fields [**51**]. However, the size of receptive fields of individual neurons increases with increasing eccentricity. This produces a non-uniform mapping of the retina and results in increasing magnification factor. The non-uniform mapping from the retina to the ocular dominance columns is schematically shown in Figure 3.1. In the fovea, one ocular dominance columns subtends approximately $10'$ (minutes of arc) of the visual field.

The information from each eye is represented continuously in the columns which correspond to that eye [**51**]. But the division of the surface of the retina amongst the ocular dominance columns is not identical in the two eyes. In other words neighbouring columns, which correspond to different eyes, do not represent identical sections of the two retinae. Specifically, the beginning of the retinal region represented in an ocular dominance column corresponding to one eye is the half-way point of the retinal region represented in the adjacent column corresponding to the other eye. With such a representation, the ocular dominance columns corresponding to the two eyes traverse the whole visual field twice by representing the surfaces of both retinae. This covering of the visual field is done in a "two

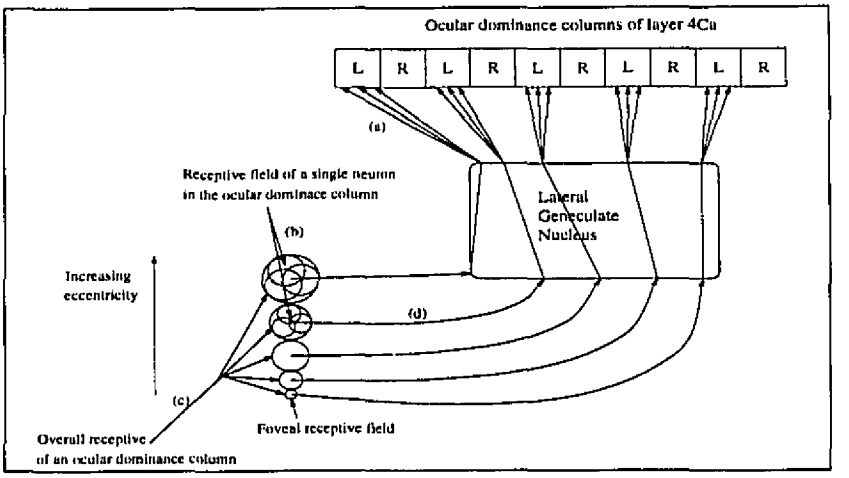


FIGURE 3.1. Schematic for Non-Uniform Mapping from the Retina to the Ocular Dominance Columns. (a) Each ocular dominance column probably receives inputs from the same number of LGN fibres. (b) But the size of the receptive fields of individual neurons increases with increasing eccentricity. (c) Therefore the overall retinal area represented in each ocular dominance column becomes larger as eccentricity increases. (d) No details are given for the mapping from the retina to the LGN.

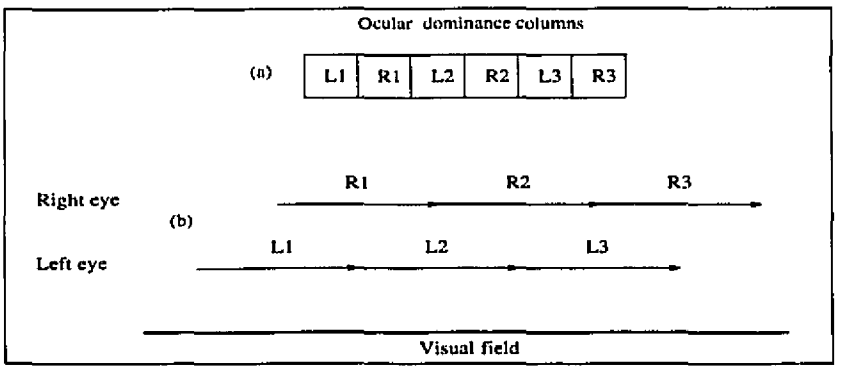


FIGURE 3.2. Schematic for the Overlap of the Retinal Areas Represented in Adjacent Ocular Dominance Columns. (a) L and R indicate the ocular dominance columns of the left and right eyes respectively. (b) Each arrow is a (1-D) illustration of the region of the retina (and of the visual field) which is represented in the corresponding column. Magnification factor is not considered.

step forward, one step backward" manner [**51**]. One ocular dominance column represents a specific section of the visual field. The next column which corresponds to the other eye starts at the mid-point of the first section and covers an area of equal size and so on. In summary, the retinal areas which are represented in the adjacent ocular dominance columns - corresponding to the two different eyes - overlap by as much as one half of the represented area. This form of representation is illustrated in Figure 3.2.

The parvocellular pathway further projects to the blob and inter-blob lattice of the layers above and below layer 4 of V1 and is believed to be involved in intensity representation as well as various stages of curve detection [2], [50], [126]. The magnocellular pathway extends from layer 4Cα to layer 4B of V1 from where it mainly projects to the thick-stripes of Visual Area 2 (V2). The thick-stripes, which are rich in cytochrome oxidase enzyme, project to Visual Area 3 (V3) and MT [28], [50].

2. Binocular Neurons

Although there is no universal computational theory for neural disparity estimation, there is a relatively broad agreement on the properties of disparity-sensitive neurons [15], [50], [91]. The investigation of binocular neurons by Poggio and colleagues [91], [92], [93], [94], [95] not only includes all neuron types observed by previous investigators [8], [13], [14], [17], [32], [50], [57], [72], [81] but also provides a more complete and accurate classification of such neurons. Poggio *et al.* have studied the response of cortical neurons to retinal disparity as well as binocular correlation of random dot patterns. The *distribution* of disparity-sensitive neurons is best described by a combination of the works of Poggio *et al.* [91], [94] and Hubel and Livingstone [50].

2.1. Ocular Balance or Dominance of Binocular Neurons. Poggio *et al.* have used bar stimuli to examine the response of cortical neurons to retinal disparity. In their experiments, they have found that some disparity-sensitive neurons respond similarly to the monocular stimulation of both eyes. Others have different responses to the individual stimulation of the two eyes. The former type of neurons illustrate ocular balance in their response to monocular stimulation and the latter type ocular imbalance or dominance to this form of stimulation. The stimulation of both eyes, or binocular summation, can result in facilitation or suppression of the monocular response. *When using elongated stimuli*, most cells are orientation selective. However, there is no evidence that orientation selectivity and disparity selectivity for line patterns are in any way related [92]. Hubel and Livingstone have also found disparity neurons sensitive to the whole range of orientations [50].

2.2. Response of Binocular Neurons to Retinal Disparity. Poggio *et al.* classify those neurons that are not sensitive to the disparity of the stimuli on the two eyes as Flat neurons and the disparity-sensitive neurons as Tuned Excitatory, Tuned Inhibitory, and Reciprocal neurons [91], [92], [93], [94], [95].

2.2.1. *Tuned Excitatory Neurons.* Tuned Excitatory (TE) neurons are ocularly balanced and do not respond to monocular stimulation of either eye. When stimulated by the same stimulus on the two retinae, their firing rate increases with respect to the resting firing rate if the stimuli are at proper disparity; the firing rate decreases with respect to the resting firing rate if the stimuli are at any other disparity. Therefore, the disparity response curve of the Tuned Excitatory cells has a clear peak at the location of the cell's preferred disparity and drops to below the resting rate for other disparities. Almost all observed TE neurons are complex cells.

In a later work Poggio *et al.* further divide the Tuned Excitatory neurons based on the magnitude of their preferred disparity and the shape of their disparity tuning curve [**91**]. Tuned Zero (T0) neurons respond to points near the horopter, with disparities less than 3′ of visual arc in the fovea. They have disparity tuning curves symmetric on both sides of the preferred disparity as shown in Figure 3.3 (a). Tuned Far (TF) and Tuned Near (TN) neurons have slightly larger preferred disparities than those of T0 neurons. In addition, the disparity tuning curves of TF and TN neurons trail towards zero disparity. Figure 3.3 (b) shows this behaviour of the disparity tuning curve for a TN cell. Near and Far refer to disparities of objects nearer and farther than the horopter respectively.

The preferred disparity range of the TE neurons depends on the eccentricity of their receptive fields. In the central 4° of the retina all observed T0 neurons have preferred disparities within $\pm 3'$ of the visual angle. Further, most TE neurons within 0.2° eccentricity have disparities less than 6′ of visual angle [**94**]. The range of preferred disparities of Tuned Excitatory neurons increases with increasing eccentricity of the corresponding foveal location [**91**], [**94**], [**95**].

2.2.2. *Tuned Inhibitory Neurons.* Tuned inhibitory (TI) neurons have properties identical to those of TE neurons with a reverse disparity tuning curve. Binocular stimulation suppresses their response at the "preferred" disparity and facilitates the response at other disparities.

2.2.3. *Near and Far (Reciprocal) Neurons.* Near (NE) and Far (FA) neurons respond to the sign of the disparity of the stimulus or equivalently to the relative location of the stimulus with respect to the horopter. Far neurons have excitatory response for uncrossed (positive) disparities which are associated with objects farther than the horopter and inhibitory response for crossed (negative) disparities associated with objects nearer than the horopter. The responses of Near neurons to uncrossed and crossed disparities is opposite

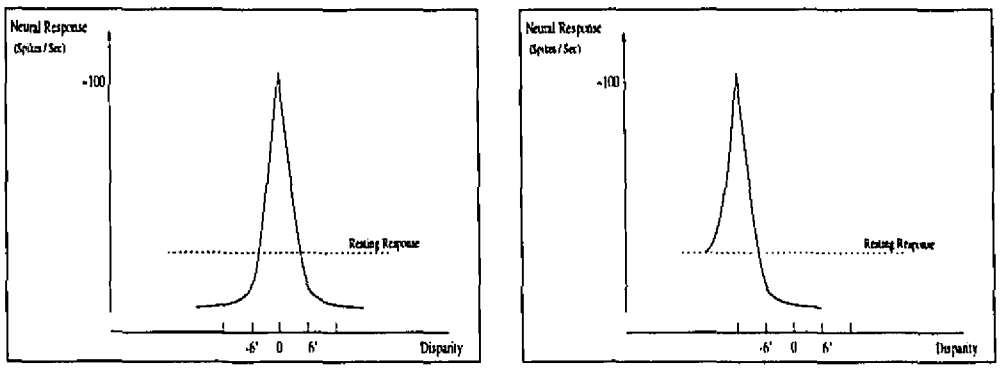


FIGURE 3.3. Response Profile of Different Tuned Excitatory (TE) Neurons. (a) Tuned Zero (TO) neuron. (b) Tuned Near (TN) neuron.

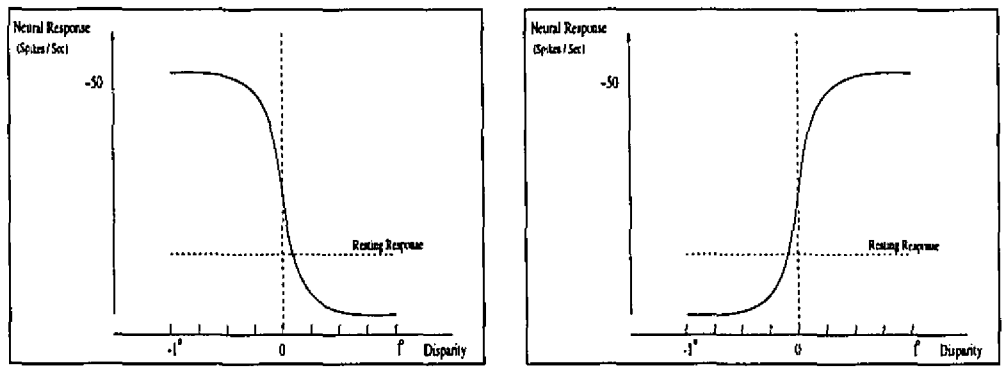


FIGURE 3.4. Response Profile of Reciprocal Neurons. (a) Near neuron. (b) Far neuron.

to that of Far neurons. The range of disparities to which reciprocal neurons respond is much larger than those of tuned neurons. Further, the drop from excitatory to inhibitory response, which occurs at zero at zero disparity, is steeper for reciprocal neurons than for tuned neurons. NE and FA neurons occur as both complex and simple cells. They also contain both ocularly balanced and unbalanced cells. The role of the reciprocal neurons may be considered complementary to those units that are unable to distinguish between crossed and uncrossed disparities [**76**]. The response profiles of reciprocal neurons are schematically represented in Figure 3.4.

2.3. Correlation Sensitivity. Poggio *et al.* also classify the binocular neurons according to their response to "correlation" or "uncorrelation" of random-dot stereogram stimuli. Binocularly uncorrelated random-dot stereograms drive the correlation sensitive

neurons to a maintained level of activity which shifts, in response to correlated images, toward facilitation or suppression as a function of positional disparity [**91**]. The resting responses of all T0 neurons are suppressed by uncorrelated random-dot stereograms. TF, TN, FA, and NE neurons respond to binocularly uncorrelated images, occupying spatially matched locations in the two eyes, mostly with activation and never with suppression. For all these neurons, the responses to uncorrelated images are usually smaller, and never larger, than the response evoked by correlated images *at the optimal excitatory disparity* [**91**].

These results may also be stated using a previous study of disparity selectivity with random-dot stereogram stimuli. The study shows that some complex cells responding to disparity of bar stimuli also respond to disparity of random-dot stimuli. Further, a number of cells responding to random-dot stereograms show no disparity selectivity for contrast bars [**92**]. *All the cells* responding to random-dot patterns are complex cells. Finally response to random-dot stereograms reflect little or no selectivity for the orientation of the binocularly visible, but monocularly hidden, figure.

2.4. Distribution of Disparity-Sensitive Neurons. The studies of disparity-sensitive neurons have involved too few neurons to give a precise indication of the distribution of these neurons in different areas of the cortex. However, based on the existing data, the ratio of disparity-sensitive to Flat neurons is approximated to be 1:1, 2:1, and 4:1 in V1, V2, and V3 respectively [**50**], [**91**]. Further, Poggio *et al.* state that these neurons occur in stripes in V2 and in clusters in V3. In V1, the disparity-sensitive neurons are found in all layers above and below layer 4C. In particular, the Tuned Excitatory neurons are mostly found in layer 4B of V1. The total number of Tuned Excitatory neurons observed equals the sum of all the other types of disparity-sensitive neurons. Equal numbers of Far and Near neurons and less Tuned Inhibitory neurons have been observed [**50**].

3. Properties of Human Stereopsis

3.1. Range of Fusible Disparities. The horopter, approximated by the Vieth Müller circle, is the locus of points whose images on the two retinae have zero positional disparity. The retinal disparity of an image point increases with increasing distance of the corresponding point in the scene from this locus. At any eccentricity, Panum's fusional area refers to the range of disparities within which the images on the two eyes can be "fused" into a single image [**77**], [**83**]. Therefore, Panum's fusional area corresponds to the range on each side of the horopter, where objects can be seen as a single object. Disparities larger the Panum's fusional limit result in seeing two images of the point rather than a single fused

image, a condition known as diplopia. Binocular visual systems compensate for diplopia by vergence or fixation on the point of interest and brings this point into sharp focus as well as zero disparity [**115**].

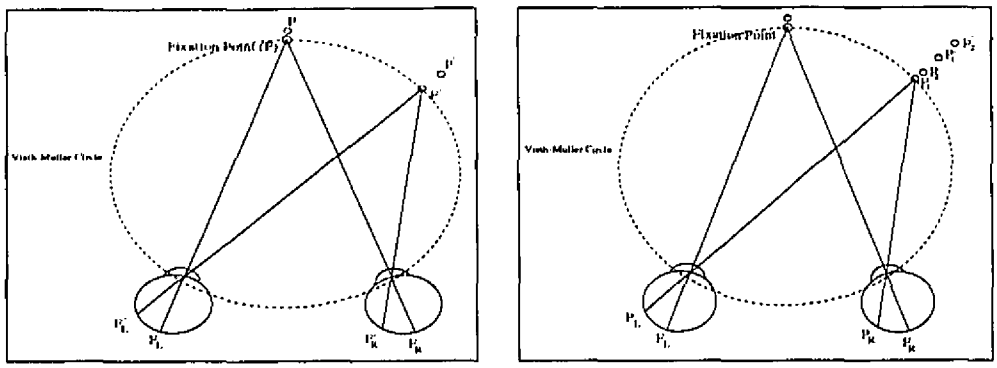
The dependence of the size of Panum's fusional area on the spatial and temporal properties of the fused stimuli is studied in [**104**], [**105**], and [**107**]. Different limits, under different measurement circumstances, have been reported for Panum's fusional area. Studies that compare many such values [**16**], [**77**], [**109**] estimate the size of Panum's fusional area in the fovea to be approximately 5′ or 6′ of visual angle, on each side of the horopter, as obtained by Ogle [**83**]. Different experiments also illustrate that the size of Panum's fusional area increases with increasing eccentricity [**16**], [**77**], [**83**]. The size of the Panum's fusional in the fovea area as well as its increase with increasing eccentricity are consistent with the range of preferred disparities of Tuned neurons.

3.2. Stereo Acuity. Stereo acuity is a measure of sensitivity to retinal disparity and is inversely proportional to stereo threshold, the minimum resolvable stereo disparity. The studies of stereo acuity, and stereopsis in general, are limited to approximately the central 10° of the retina since at higher eccentricities even the determination of fusion and diplopia becomes difficult [**60**], [**77**].

The experiments in [**31**], [**97**], [**106**], and [**113**] show that stereo acuity decreases with increasing distance from the fovea. In other words, as the image moves away from the fovea *along the horopter*, the minimum resolvable *absolute* disparity increases. Stereo threshold also increases with increasing absolute disparity or distance from the horopter. Equivalently, at a given eccentricity, as images move *away from the horopter* the minimum detectable *relative* disparity increases [**16**], [**106**]. The two variations of stereo threshold are schematically illustrated in Figure 3.5. Blakemore [**16**] specifically describes the latter relationship as exponential. Schor [**106**] mentions that the fall in stereo acuity with increasing distance *from* the horopter is more than the fall with increasing eccentricity *along* the horopter.

Using the analogy of disparity-sensitive neurons, stereo acuity with respect to the horopter can be related to the function of the Tuned neurons. As explained in Sections 2.2.1 and 2.2.2, at any eccentricity Tuned neurons respond to small disparities associated with points near horopter. Stereo acuity would then be analogous to the resolution of the response curve of such neurons.

It is also important to mention another concept, referred to as "stereo positional acuity" in [**125**]. Stereo positional or spatial acuity refers to the ability to perceive *multiple* spatial changes in depth or stereo disparity, as opposed to discriminating between two surfaces at



(a) (b)

FIGURE 3.5. Variations in Stereo Acuity and Stereo Threshold (schematics not to scale.) (a) The disparity of a point, P, is compared to that of the horopter or zero disparity. Therefore, the relative disparity between the two equals the absolute disparity of the test point. With increasing eccentricity of the point, the minimum detectable absolute (or relative to zero) disparity increases. (b) the disparities of *two* test points, 1 and 2, are under consideration. At a given eccentricity, as the absolute disparity (or distance with respect to the horopter) of one point increases, the minimum detectable disparity difference between the two points also increases.

different depths used in the definition of stereo acuity. The experiments in [**121**] illustrate that the maximum detectable disparity and even stereo fusion decrease with increasing spatial frequency of disparity change. Specifically, for spatial frequencies near or greater than one cycle per twenty minutes of visual angle, 20′, stereo performance deteriorates significantly [**121**].

3.3. Stimulus Contrast, Inter-ocular Correlation, and Opposite Contrast Stimuli. Inter-ocular correlation (IOC) is defined as the cross-correlation function between the left and right images of a stereo image pair [**119**]. The experiments in [**27**] show that there is an inverse square relationship between inter-ocular correlation and stimulus contrast in stereo perception. This means that to maintain stereopsis, stimulus contrast has to be increased by the square of any given decrease in IOC.

The studies of [**26**] indicate that for images with complexities, or random-dot stereograms with densities higher than a certain minimum, human subjects are unable to fuse opposite contrast stereo image pairs. Also Poggio *et al.* [**91**] mention that disparity-sensitive neurons, and particularly those sensitive to binocular correlation do not respond to anti-correlated stimulation of the two eyes.

4. Human Stereopsis: Absolute Surface Description Versus Relative Depth Perception

As mentioned in Chapter 2, one can obtain the complete depth profile of a scene using horizontal disparity information about the scene and knowledge of imaging geometry or vertical disparities. In biological stereopsis, one may assume that the oculomotor system provides the required information about imaging geometry; one can also argue the existence of channels responsive to vertical disparity although, to our knowledge, no indication of such channels has been observed. Therefore, in most theories of stereopsis perception of the third dimension is assumed to involve computing stereo disparity, depth, and shape in that order.

Stevens and collaborators take a different approach to three-dimensional perception, founded on the properties of human stereopsis rather than geometrical plausibility. Psychophysical experiments in [**116**] illustrate that human stereopsis is highly insensitive to constant gradients of disparity, or equivalently to constant rates of change in depth. The experiments in [**116**] also illustrate sensitivity to the non-zero second derivative of disparity, or equivalently to surface curvature or depth discontinuities. Based on these observations, Stevens and Brookes conclude that stereopsis is a source of information about surfaces whose depths change abruptly or at least at a non-constant rate, not about absolute depth or even linear depth profiles [**115**], [**117**], [**116**]. Also, experiments with random-dot stereograms in [**21**] show that identical disparity profiles can result in perception of different shapes when embedded within different background disparities. Grimson [**42**] provides another example where two surfaces of equal depth, when placed next to neighbours of different depth profiles, are perceived as having different distances. In other words, it is the disparity profiles of both the surface and its background, or the relationship between the two, that influence the perceived shape rather than the disparity of the surface alone. The experiments of [**38**] and [**39**] provide further evidence that disparity discontinuities are used in stereoscopic processing. Finally, a quantitative value describing the "salience" of the disparity of each image feature is defined in [**78**]. Disparity salience is a function of the *differences* between the disparity of the feature and those of its neighbours. The reason for such a definition is that neighbouring disparities influence the perceived depth of the test feature [**78**].

Such observations lead to a different view of human stereo vision in which stereo disparity is not a mean for measuring absolute depth; stereopsis, along with other depth cues, results in *information about* the three dimensional *shape* of the scene. The information from *all* depth cues then, collectively result in more precise knowledge of depth. Different sources of depth information, such as stereopsis, occlusion, motion parallax, or shading may

of course, provide consistent or rival cues resulting in different perceptions in different observers. Stevens and colleagues show that in many occasions, other depth cues are dominant to stereo disparity [**115**], [**117**], [**118**].

Whatever the integration strategy, it is evident from these experiments that stereopsis is not the sole or even the dominant mean of depth perception. McKee *et al.* have studied the precision of stereopsis and have concluded that distance perception with stereopsis is imprecise [**76**]. With such evidence, the role of stereopsis seems to be detecting those places where depth changes at a non constant rate, or equivalently distinguishing the boundaries between regions of constant (including zero) depth gradient. Such a role for stereopsis, rather than its more tradi nal role as a mean for exact surface reconstruction, constitutes the approach of this thesis to exploiting stereo disparity information.

CHAPTER 4

Autocorrelation, Cepstrum, and Binocular Disparity

For all points in a stereo image pair, except those affected by occlusion, there is a corresponding point in the second image. Binocular disparity, by definition, is the shift between the locations of the corresponding points in the image pair. Therefore, the images of a stereo pair constitute signals of which one is an identical and shifted version of the other. In other words, one image of the stereo pair is a *spatial echo* of the other image. The shift of course achieves different values at different parts of the image depending on the distance of the corresponding surface from the imaging device. This chapter studies autocorrelation and the cepstrum as tools for detecting echos and estimating the shift that generates an echo from an original signal. In the context of stereo vision, the shift or delay between the two signals is referred to as the disparity of the image pair. Therefore the terms echo shift, delay, or disparity are used interchangeably throughout the chapter.

The chapter starts with autocorrelation which is a more familiar concept. Then the cepstrum is motivated as a response to a shortcoming of autocorrelation. The cepstrum is first introduced as a modification to autocorrelation and later studied as a distinct concept. Throughout the study of the cepstrum, autocorrelation is used to aid in visualizing the discussed properties.

1. Autocorrelation

The autocorrelation function, $R_g(\tau)$, of a real signal $g(x)$ is defined as

$$R_g(\tau) = g(x) * g(-x) = \int_{-\infty}^{+\infty} g(x)g(x-\tau)dx \tag{4.1}$$

where $*$ denotes the convolution of two signals. As shown in Appendix A, the Fourier transform of the autocorrelation function is the power spectrum of the signal. Equivalently if the Fourier transform of $g(x)$ is denoted by $G(f)$,

$$R_g(\tau) \rightleftharpoons \mid G(f) \mid^2 \tag{4.2}$$

The power spectrum of a real signal is an even function of frequency. Therefore, both forward and inverse Fourier transforms of the power spectrum result in the autocorrelation function, except for a scaling factor.

In Equation 4.1 the time lag τ plays the role of a scanning or searching parameter and the autocorrelation function, $R_g(\tau)$, provides a measure of similarity between the waveforms of the functions $g(t)$ and $g(t-\tau)$ [48]. Equivalently, the autocorrelation function can be used as a similarity indicator or correspondence detector between various sections of a signal. The higher the correspondence between the original signal and its shifted version, the greater the value of autocorrelation is at that shift.

2. Autocorrelation in the Presence of Echo

2.1. Autocorrelation of a Signal Containing Echo. Echo can be defined as a shifted and possibly scaled version of the original signal. Therefore it can be described using convolution with a shifted delta function. An original signal $s(x)$ with an added echo of delay D and scaling factor 1 can be represented as

$$g(x) = s(x) + s(x-D) = s(x) * (\delta(x) + \delta(x-D)) \tag{4.3}$$

Assuming real signals, the autocorrelation function of $g(x)$ is then

$$\begin{aligned} R_g(\tau) &= \int_{-\infty}^{+\infty} g(x)g(x-\tau)dx \\ &= \int_{-\infty}^{+\infty} (s(x)+s(x-D))(s(x-\tau)+s(x-D-\tau))dx \\ &= \int_{-\infty}^{+\infty} s(x)s(x-\tau)dx + \int_{-\infty}^{+\infty} s(x)s(x-D-\tau)dx \\ &\quad + \int_{-\infty}^{+\infty} s(x-D)s(x-\tau)dx + \int_{-\infty}^{+\infty} s(x-D)s(x-D-\tau)dx \\ &= 2R_s(\tau) + R_s(\tau+D) + R_s(\tau-D) \end{aligned} \tag{4.4}$$

or equivalently

$$R_g(\tau) = R_s(\tau) * (2\delta(x) + \delta(x+D) + \delta(x-D)) \tag{4.5}$$

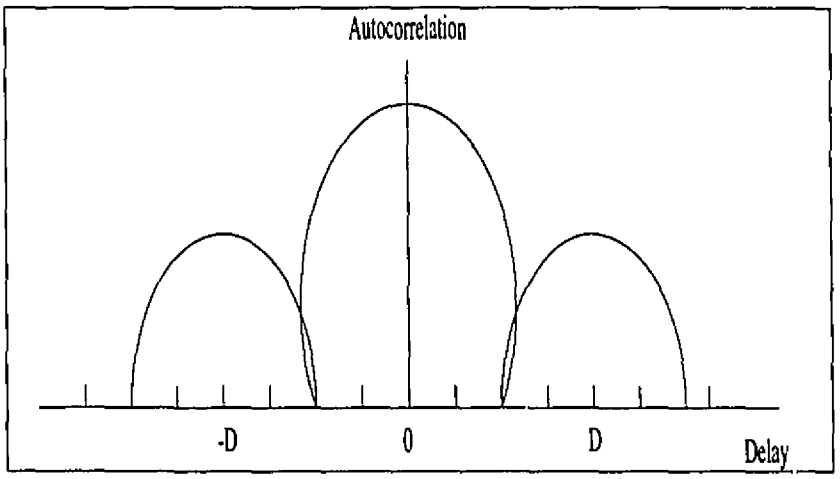


FIGURE 4.1. The Autocorrelation Function of a Signal Containing an Original and its Echo (schematic). The autocorrelation function of the overall signal consists of the autocorrelation of the original and its shifted version.

As illustrated in Equations 4.4 and 4.5, the autocorrelation function of a signal containing an original and an echo equals the sum of twice the autocorrelation of the original and two copies of its shifted version. The shift equals that of the echo. Figure 4.1 schematically represents the autocorrelation function of Equation 4.5.

2.2. Obtaining the Echo Shift from the Autocorrelation Function. The autocorrelation function of the overall signal has contributions from two factors: the original signal and the *presence* of a shifted version of this signal. $R_s(\tau)$ is due to and completely determined by the original signal; *repetition* of $R_s(\tau)$ is due to the presence of the echo. The two contributions are *interconnected* by the fact that it is *the autocorrelation of the original that repeats.* In Equation 4.5, $R_s(\tau)$ and the shifted impulses represent the "original signal" and "shift" components of the autocorrelation. The convolution provides the interconnection between them.

The shift component, represented as shifted impulses in the autocorrelation function, appears as a cosine wave in the frequency domain. In other words, the spectrum of the original signal is multiplied by a periodic (cosine) wave whose frequency, in the frequency domain, equals the shift of the echo. It is important to note that this periodic component does *not* exist independently from the spectrum of the original signal but as a multiplicative factor. The magnitude of this multiplicative factor is independent of the spectrum of the original signal.

In measuring the delay generating the echo, one is interested in the shift components of the autocorrelation function or the shifted impulses. Under such circumstances, the autocorrelation of the original signal is an unwanted component acting as "noise". To facilitate obtaining the shift component, one can treat this convolved noise in two ways: suppress its relative energy or convert it from non-additive to additive noise.

2.2.1. *Suppressing the Signal Component.* For a given signal energy, if the spectrum of the original signal is approximately constant or white over the frequency band of interest, the sinusoids representing the shift component reside over a large range of frequencies. On the other hand, if the spectrum of the original signal has large magnitude components in a limited band of frequencies, the multiplicative sinusoids exist over a smaller range. Equivalently, for such signals there are large original-signal components and small shift components. Suppressing the original-signal component relative to the shift component increases the saliency of the latter.

Application of a compression function to the power spectrum suppresses the larger original-signal components more than the smaller shift components. In other words, the compression function acts as a "spectrum whitener". Examples of compression functions include the fourth root, the tan function, and the logarithmic transformation [**86**]. Applying the logarithm to power spectrum before it undergoes Fourier transformation, is equivalent to finding the autocorrelation of the filtered version of the original signal [**86**]. The overall transformation can be thought of as autocorrelation with an adaptive nonlinear pre-filter which has the property of being compressive in the frequency domain. It tends to make the power spectrum more uniform by reducing the contribution of narrow band signals while leaving the broadband signals relatively unaltered [**85**]. Figure 4.2 is an illustration of the result of suppressing the original-signal component of the autocorrelation function of Figure 4.1.

2.2.2. *Separating the Shift Component From the Signal Component.* The interconnection between the original-signal and the shift components is via multiplication in the frequency domain, converted to convolution after the Fourier transformation. If the multiplication in the frequency domain is somehow replaced by an addition, the linearity of the Fourier transformation results in a sum rather than the convolution of the two components in the time or spatial domain. In this way, there are two *added* terms after the Fourier transformation: a term due to the Fourier transform of the spectrum of the original and

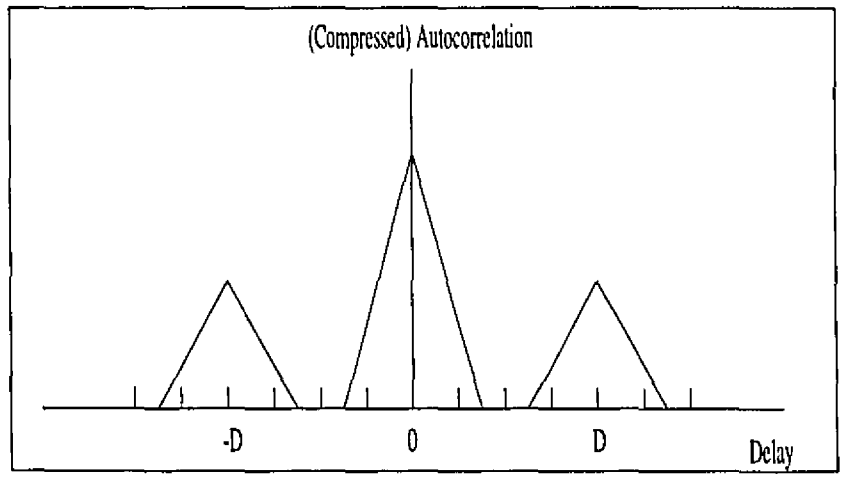


FIGURE 4.2. The Autocorrelation Function of a Signal Containing an Original and Echo with Suppressed Original-Signal Component (schematic).

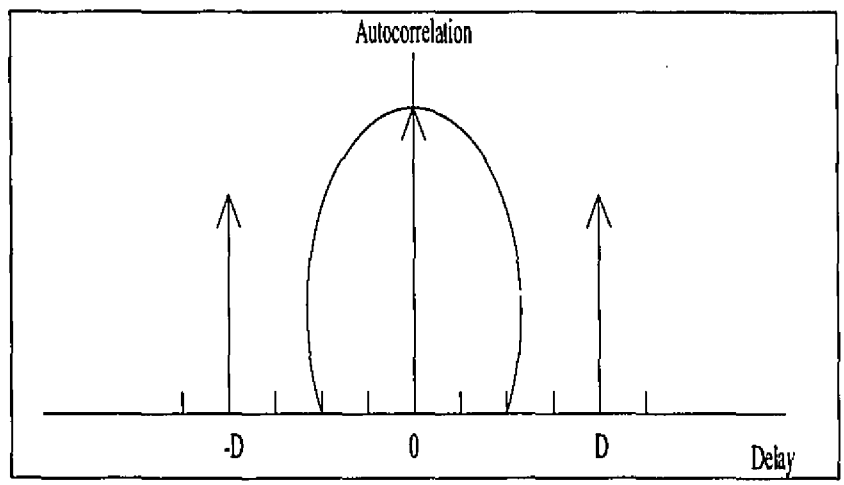


FIGURE 4.3. The Autocorrelation Function of a Signal Containing an Original and Echo with Separated Original Signal Component (schematic).

another due to the echo. This conversion of non-additive to additive noise is shown in Figure 4.3.

Conversion from multiplication to addition is most easily performed by taking the logarithm of the power spectrum. In other words, a logarithmic transformation converts the multiplicative periodic components in the spectrum to additive periodic components in the log spectrum.

With such a property, in addition to suppressing the original-signal component of autocorrelation, the logarithmic transformation achieves deconvolution of the two components.

This effect is similar to the deconvolution by phase correlation suggested in [86]. Logarithmic deconvolution is advantageous to phase correlation because of its inherent compressive effect, described in the previous section.

3. The Cepstrum

The idea of the logarithmic transformation of the power spectrum for enhancing the effects of echos in autocorrelation is in fact not new in signal processing. In 1963, Bogert *et al.* introduced the cepstrum and cepstral analysis in [18] as a tool for echo detection. Shortly afterwards, Oppenheim *et al.* extended the notion of cepstrum to generalized nonlinear filtering of convolved signals [88]. Oppenheim *et al.* [88] refer to the definition of the cepstrum given in [18] as the power cepstrum of a signal. As defined in [18], the power cepstrum of a signal is the power spectrum of the logarithm of the power spectrum of the signal. Since then, various sources have offered other definitions for the power cepstrum. In particular, considering that signal processing is normally concerned with discrete signals, the definition of the power cepstrum has been extended to the z-transform of signals. Childers *et al.* [25] define the power cepstrum of a signal to be the square of the inverse z-transform of the logarithm of the magnitude squared of the z-transform of the signal. Oppenheim and Schafer [87] evaluate the z-transform on the unit circle. In this manner, they define the power or real cepstrum of a signal as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of the signal.

All the above definitions of the cepstrum are to a large extent equivalent. Considering the logarithmic transformation, the squaring involved in some definitions (taking the logarithm of the power spectrum versus the logarithm of the magnitude of the Fourier Transform) only results in a scaling of the final value. Throughout this thesis, the definition of the power cepstrum given in [87] is used and the Fourier transform is approximated by the discrete Fourier transform. Using the notation of the continuous Fourier transform,

$$g_{cep}(x) = \int_{-\infty}^{\infty} \log \mid G(f) \mid e^{j\pi f x} df \tag{4.6}$$

Note that taking the Fourier transform maps information back into the same domain as the that of the original signal. Bogert *et al.* [18] refer to the domain of the power cepstrum as the *quefrency* domain.

If the signal $g(x)$ is formed through the convolution of two signals $s(x)$ and $f(x)$, transformation to the frequency domain has the property of changing the convolution to

multiplication; taking the logarithm results in the addition of the log spectra of the two signals. Equivalently, if

$$g(x) = s(x) * f(x) \tag{4.7}$$

then

$$| G(f) | = | S(f) | . | F(f) | \tag{4.8}$$

or

$$\log | G(f) | = \log | S(f) | + \log | F(f) | \tag{4.9}$$

Taking the inverse Fourier transform of Equation 4.9 gives the power cepstrum of $g(x)$ to be

$$g_{cep}(x) = s_{cep}(x) + f_{cep}(x) \tag{4.10}$$

Appendix B shows that if the signal $g(x)$ consists of an original signal $s(x)$ and its shifted version or echo, then the echo appears as ripples in the log spectrum of $g(x)$ and as impulses in $g_{cep}(x)$ as in Equations 4.11 and 4.12. The delay of the echo determines the frequency of the ripple and the location of the impulses. As seen in Appendix B, the multiplicity of impulses is due to multiple convolutions corresponding to the power series expansion of the logarithmic expression.

$$\log | G(f) | = \log | S(f) | + \sum_{m=1}^{\infty} (-1)^{m+1} \frac{(ae^{-j2\pi fD})^m}{m} \tag{4.11}$$

$$g_{cep}(x) = s_{cep}(x) + \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\delta(x - mD)}{m} \tag{4.12}$$

Figure 4.4 shows a one-dimensional signal as a function of time and its delayed version. The resultant signal is the sum of the original and its echo . The ripples in the log spectrum are seen in Figure 4.4 (d). Figures 4.4 (f) and (g) show the impulses in the power cepstrum of the resultant signal at the quefrency equal to the delay of the echo.

The concept of cepstral analysis can be easily extended to two-dimensional signals and the two-dimensional Fourier transform. For two-dimensional signals, the power cepstrum

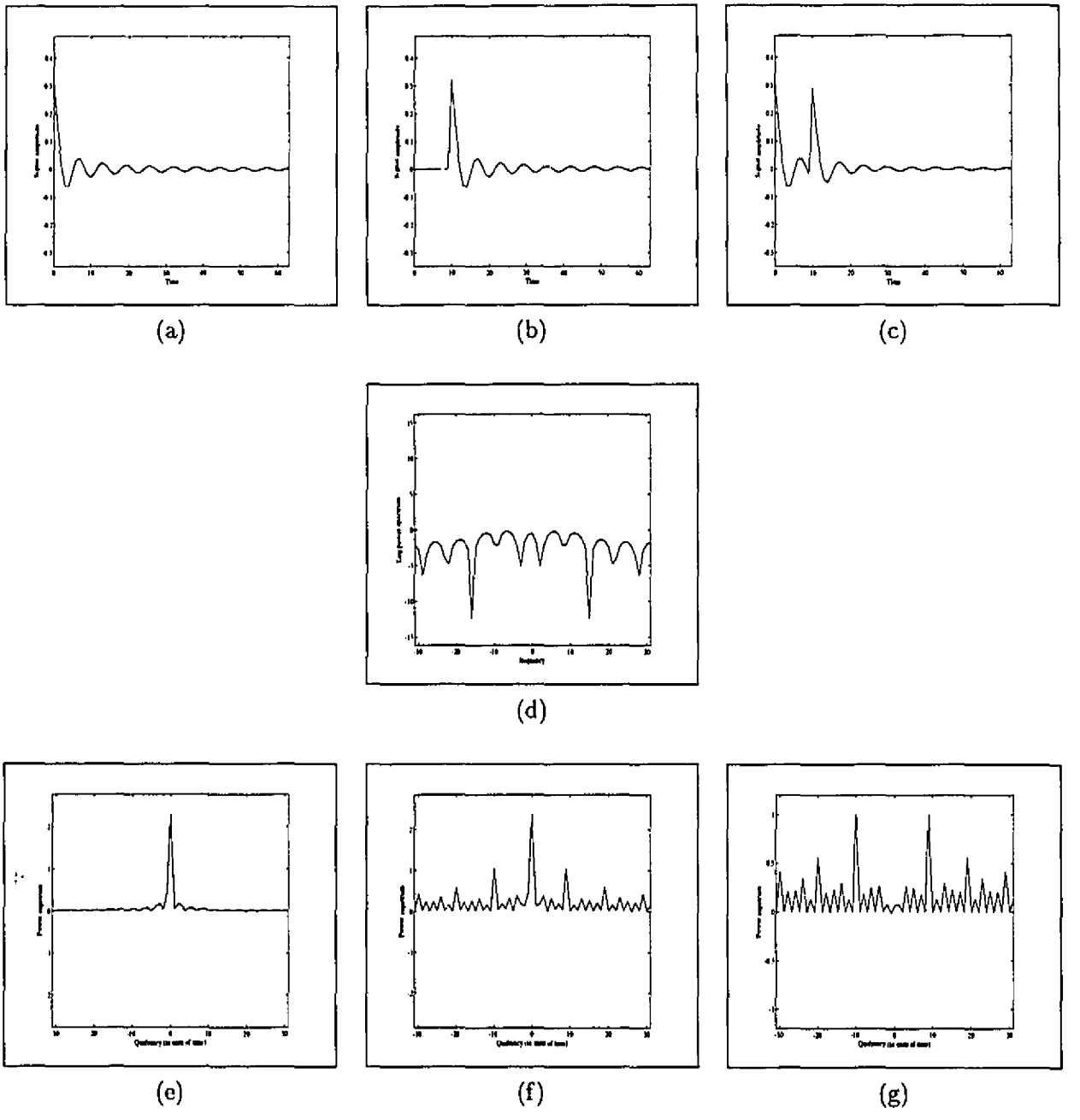


FIGURE 4.4. Cepstral Analysis of a One-Dimensional Signal with an Added Echo. (a) Original signal. (b) Echo (delayed signal) with a delay of 10 samples. (c) Resultant signal from addition of the original and echo. (d) Log spectrum of the resultant signal. (e) The power cepstrum of the original signal. (f) The power cepstrum of the resultant signal. (f) The power cepstrum of the resultant signal with the power cepstrum of the original signal removed.

of a signal containing an original and an echo contains an impulse at the location specified by echo delays in both directions.

4. Echo Resulting from an Array of Sensors: An Alternative Data Representation

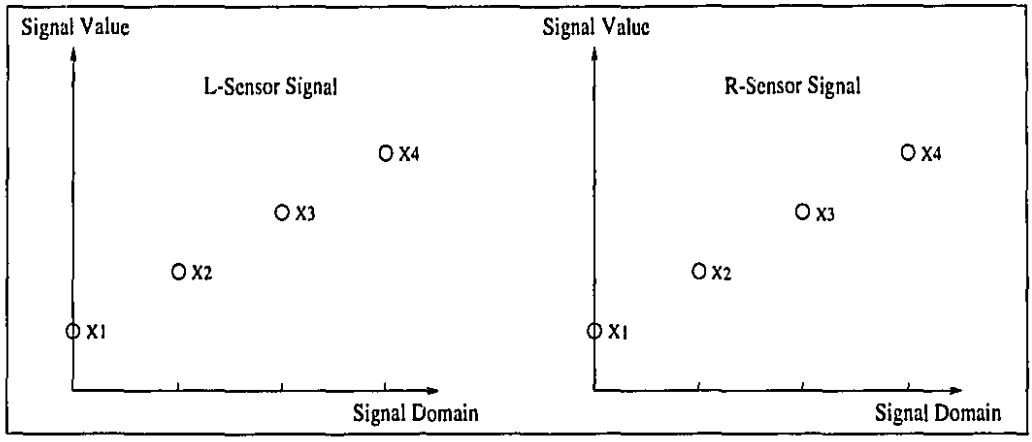
The signal received by each sensor in an array of sensors is a spatially or temporally shifted version of the signals at the other sensors in the array. Therefore, with an array of sensors it is possible to obtain many copies of a signal, each a shifted version of the others. The separate availability of multiple copies of the signal suggests two different ways for obtaining the shift between two of the received signals, using a correlation based algorithm. First, it is possible to add the two signals and estimate the echo shift in the resultant signal using the methods described above. Second, one can concatenate the two signals and form a signal chain whose length is twice the length of each individual signal. The resultant signal still includes an original signal and its shifted version, but the shift is different from that of the initial two signals. When the original and echo are added, the delay of the resultant signal simply equals the delay of the echo. When the two signals are placed next to each other, the total shift of the resultant signal equals the vectorial sum of the shift in the echo and the shift due to placing one signal next to the other. The latter of course equals the length of one of the two signals, assuming that both signals have the same length.

The effect of concatenating the original signal and the echo can be seen in Figure 4.5 for the same signal as Figure 4.4. Figures 4.5 (d) , (f) , and (g) show the higher frequency of ripples in the log spectrum and the shifted location of the impulse in the power cepstrum.

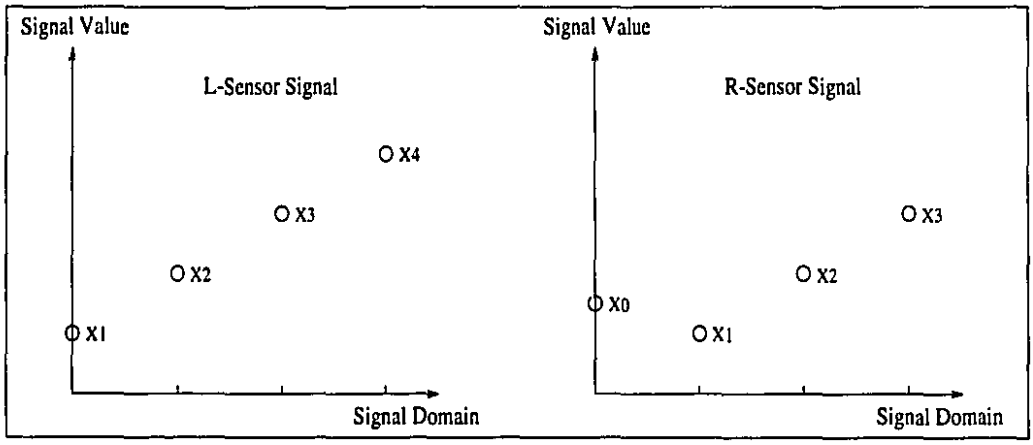
5. Data Dependence of Analysis using Autocorrelation and the Cepstrum

The performance of the cepstrum in echo detection is known to be extremely data dependent [**25**]. The data dependence of the cepstrum, as well as autocorrelation, can be explained using their definitions and that of an echo. This section uses the autocorrelation function, which is the simpler concept of the two, to explain the reasons for the dependence on data. The explanation is then extended from the autocorrelation to the cepstrum. It is assumed that the original and echo signals are obtained from distinct sensors. As explained in Section 4, when the original and echo signals are separately available, the echo delay can be estimated by adding or concatenating the two signals.

5.1. Consequences of Finite Signal Length. In deriving the concept of echo detection using autocorrelation or the cepstrum, it is inherently assumed that both the original



(a)



(b)

FIGURE 4.6. Number of Corresponding Samples in Finite Length Signals. The signals are taken from two distinct sensors, referred to by L and R. X0 to X4 illustrate the samples that are taken from the two sensors. There is a corresponding sample for Xn from one sensor if Xn also appears in the signal from the other sensor. (a) With zero delay all samples of the signal from each sensor have a corresponding sample in the signal from the other sensor. (b) The delay between the two signals is one sample. With a non-zero delay some samples from each sensor do not have a corresponding sample in the signal from the other sensor.

signal and the echo contain identical information. For delays other than zero, this assumption is equivalent to assuming that signals are infinitely long. Of course, in signal processing applications one uses signals of finite length. Therefore, for all non-zero disparities there is a segment at the beginning or end of each signal for which no corresponding segment in the other signal exists. This phenomenon is schematically illustrated in Figure 4.6.

5.1.1. *Autocorrelation Peak and Finite Signal Length.* As stated in Section 1, the autocorrelation function of a signal can be obtained by convolving the signal with its shifted version. It is also known that performing convolution by the multiplication of Fourier transforms results in circular convolution [**24**]. Equivalently, the value of the autocorrelation function at any shift can be obtained by summing the products of the samples of the signal and corresponding samples of its shifted version. Circular convolution should be taken into consideration while obtaining the products.

Section 4 stated that signals obtained from distinct sensors can be represented by addition or concatenation. Figure 4.7 (a) contains the resultant signal obtained from *sample-by-sample adding* of the two individual signals of Figure 4.6 (a). Figure 4.7 (b) contains a similar resultant signal for the two signals of Figure 4.6 (b). Figures 4.6 (c) and (d) contain the shifted versions of the signals in Figures 4.6 (a) and (b) respectively. The shift generating each of the signals in (c) and (d), equals the delay that exists between the generating signals from the two sensors.

Figures 4.8 illustrates the same information as Figure 4.7 for signals that are *placed next to each other.* As explained in Section 4, when an original and an echo are concatenated, the resultant signal includes an echo whose delay is different from the original. The delay present in the resultant signal equals the vectorial sum of the initial disparity and the length of one of the initial signals. This resultant delay is the shift which generates the signals of Figures 4.8 (c) and (d) from those in (a) and (b) respectively.

In the remainder of this section, the term *correct shift* is used to refer to the shift which corresponds to the disparity between the original signal and echo while taking the specific data representation into consideration. Therefore, for data representation using addition of signals, the *correct shift* equals the disparity between the original and the echo. For data representation using concatenation, it equals the disparity between the original and the echo plus the length of one of the two.

Consider Figures 4.7 and 4.8 showing the addition and concatenation representations respectively. From these figures one can obtain the value of the autocorrelation function of each resultant signal at the correct shift. This value simply equals the sum of the sample-by-sample products of each resultant signal with its shifted version. As seen in the figures, when the delay between the original and echo signals is zero, at the correct shift, all the samples of the resultant signal are multiplied by identical samples in its shifted version. On the other hand, with non-zero disparity at the correct shift, only a fraction of the samples align with identical samples in the shifted signal. This fraction equals the number of samples

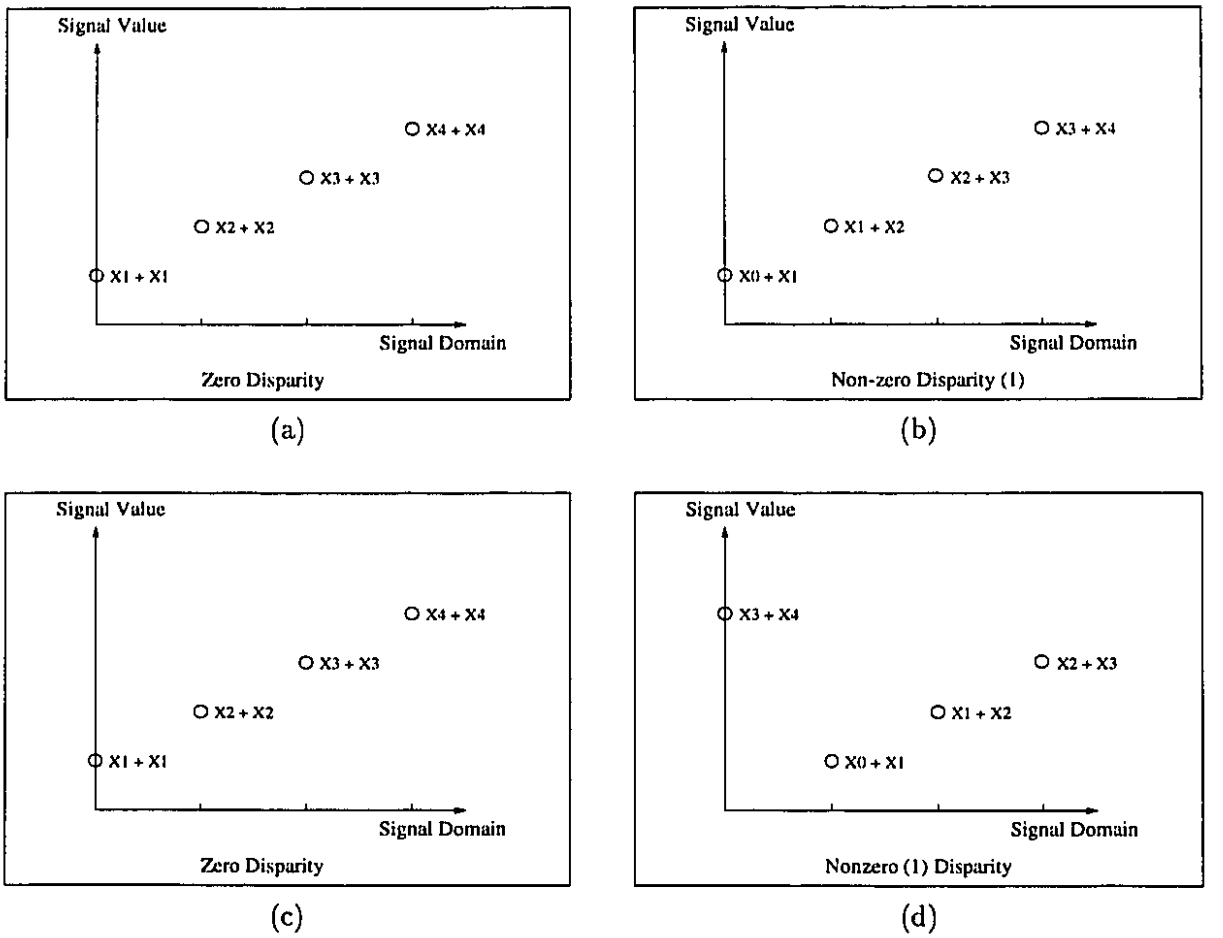


FIGURE 4.7. Autocorrelation of the Signal Resulting from Adding the Original and Echo Signals. The original signals are shown in Figure 4.6. (a) The resultant signal from adding the two signals in Figure 4.6 (a). (b) The resultant signal from adding the two signals in Figure 4.6 (b). (c) The shifted version of signal in (a). The shift equals the delay between the two signals in Figure 4.6 (a) which is zero. (d) The shifted version of signal in (b). The shift equals the delay between the two signals in Figure 4.6 (a) which is non-zero (one). In generating the shifted signals wrap-around has been taken into consideration.

that are in common between the original and echo signals. Assume that each of the original and echo signals are W samples long. For zero disparity and representation by addition, the value of the autocorrelation function at the correct shift includes W products of identical samples; It includes $2W$ products of identical samples for representation by concatenation. For any non-zero disparity d , the value of the autocorrelation function at the correct shift includes $W - d$ products of identical samples in both representations.

If the signal is uncorrelated or even has a decreasing autocorrelation function, the decrease in the number of contributing samples results in a decrease in the magnitude of the

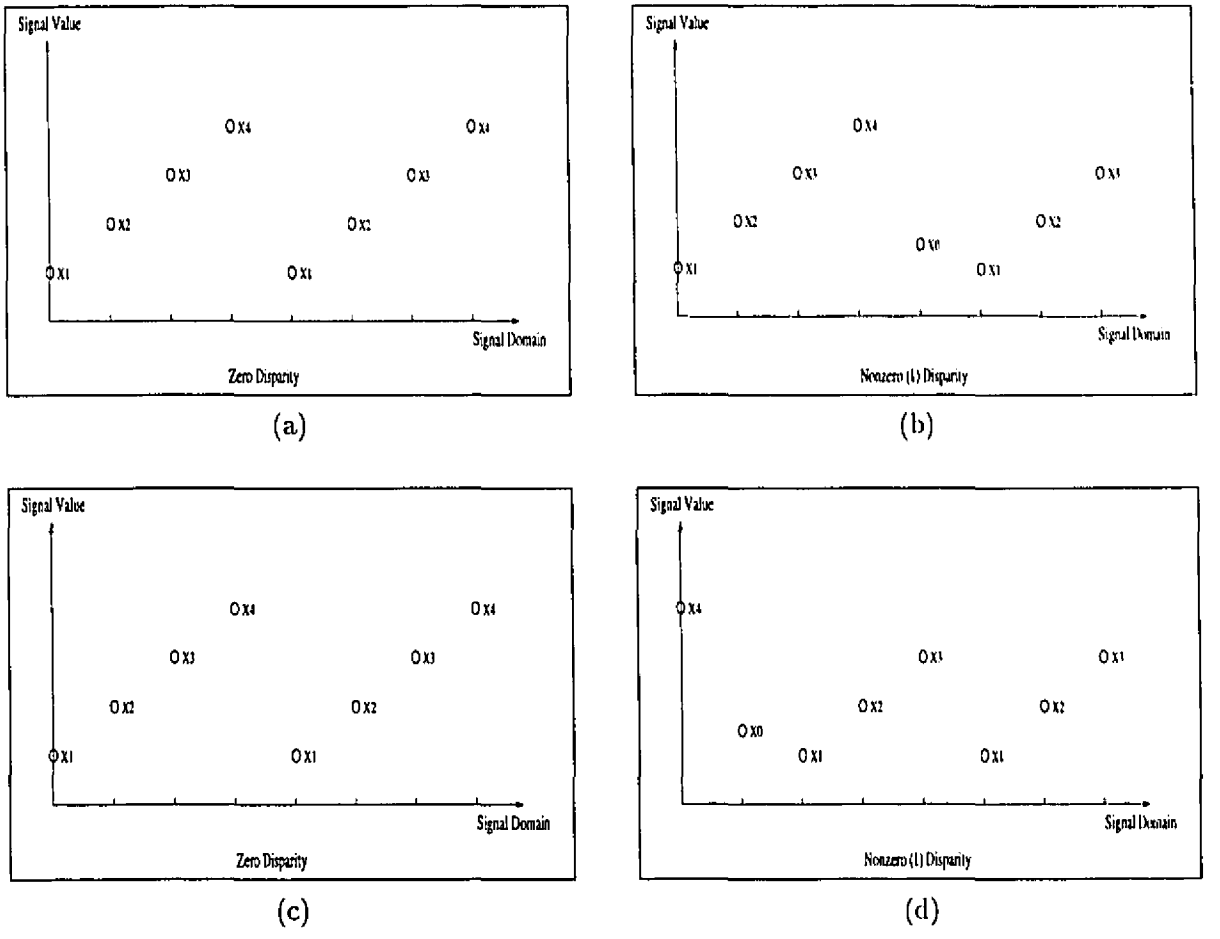


FIGURE 4.8. Autocorrelation of the Signal Resulting from Concatenating the Original and Echo Signals. The original signals are shown in Figure 4.6. (a) The resultant signal from concatenating the two signals in Figure 4.6 (a). (b) The resultant signal from concatenating the two signals in Figure 4.6 (b). (c) The shifted version of signal in (a). The shift equals the delay between the two signals in Figure 4.6 (a) (zero) plus the length of each individual signal (four). (d) The shifted version of signal in (b). The shift equals the delay between the two signals in Figure 4.6 (a) (one) plus the length of each individual signal (four). In generating the shifted signals wrap-around has been taken into consideration.

autocorrelation peak. Therefore, although the autocorrelation function always has a maximum at the shift corresponding to the correct disparity, the magnitude of its peak decreases with increasing disparity. In particular, the drop in magnitude from zero disparity to a disparity of one is very large for the concatenated representation. This sudden decrease is due to the artificially introduced shift. For uncorrelated data, the ratio of the autocorrelation peak at any disparity d to that at disparity 0 can be approximated as:

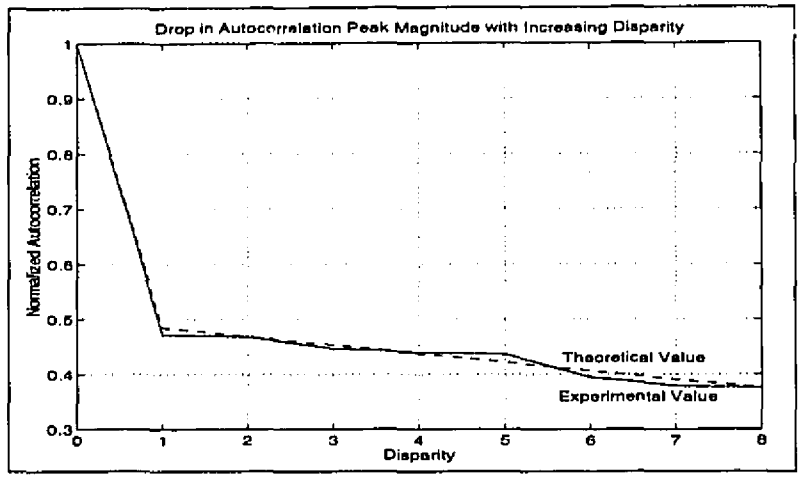


FIGURE 4.9. Decrease in Autocorrelation Peak Magnitude with Increasing Disparity. The length of the original signal is 32 samples and signals are represented by concatenation.

$$\frac{Peak_{autocorr}(d)}{Peak_{autocorr}(0)} = \begin{cases} 1 & d = 0 \\ \frac{W-d}{W} & d \neq 0 \end{cases} \tag{4.13}$$

when the original and echo signals are added, and

$$\frac{Peak_{autocorr}(d)}{Peak_{autocorr}(0)} = \begin{cases} 1 & d = 0 \\ \frac{W-d}{2W} & d \neq 0 \end{cases} \tag{4.14}$$

when the original and echo signals are placed next to each other.

Figure 4.9 illustrates the theoretical and experimental decrease in the magnitude of the autocorrelation peak when data representation using concatenation is used. The signal is a zero mean random signal.

5.1.2. *The Cepstrum and Finite Signal Length.* To ensure that the logarithm in Equation 4.6 is not undefined when using the cepstrum for echo detection, one can add a constant - 1 for example - to the argument of the *log* function. Using the Taylor Series expansion

$$\log(1+z) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{z^m}{m} \tag{4.15}$$

which is valid for $\mid z \mid \leq 1$ and $z \neq -1$, we obtain

$$g_{cep}(x) = \int_{-\infty}^{\infty} \sum_{m=1}^{\infty} (-1)^{m+1} \frac{|\, G(f)\,|^m}{m} e^{j\pi f x} df \tag{4.16}$$

Using the definition of autocorrelation and the multiplication-convolution duality property of Fourier transformation, the cepstrum can be expressed as

$$g_{cep}(x) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{(R_g(\tau))^{m*}}{m} \tag{4.17}$$

where $(\bullet)^{m*}$ indicates $m-1$ convolutions of the argument with itself. The exact relationship between the peak magnitudes of the cepstrum and autocorrelation is dependent on the correlation properties of the signal. However, as shown in Appendix C for uncorrelated signals the ratio of the cepstral peak magnitude for any disparity d to that of disparity zero can be expressed as follows:

$$\frac{Peak_{ceps}(d)}{Peak_{ceps}(0)} = \begin{cases} 1 & d = 0 \\ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{\sum_{i=0}^{\lfloor \frac{m-1}{2} \rfloor} \left(\frac{W-d}{W}\right)^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!}}{\sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} \left(\frac{W-d}{W}\right)^{2i} \frac{m!}{(m-2i)!i!i!}} & d \neq 0 \end{cases} \tag{4.18}$$

when the original and echo signals are added, and

$$\frac{Peak_{ceps}(d)}{Peak_{ceps}(0)} = \begin{cases} 1 & d = 0 \\ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{\sum_{i=0}^{\lfloor \frac{m-1}{2} \rfloor} \left(\frac{W-d}{2W}\right)^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!}}{\sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} \left(\frac{W-d}{2W}\right)^{2i} \frac{m!}{(m-2i)!i!i!}} & d \neq 0 \end{cases} \tag{4.19}$$

when the original and echo signals are placed next to each other. $\lfloor \bullet \rfloor$ indicates the integer part of the argument.

Figure 4.10 illustrates the theoretical and experimental decrease in the magnitude of the cepstral peak when data representation using concatenation is used. The first ten terms in Equation 4.19 have been used for obtaining the theoretical value. The signal is a zero-mean random signal.

5.2. Inherent Data Correlation and the False Disparity Problem. For highly uncorrelated data, the autocorrelation function has a clear peak at the shift corresponding to the disparity between the original and echo signals. The cepstrum eliminates the contribution of the signal itself and represents the disparity by an impulse at the corresponding shift. Therefore, the autocorrelation peak and cepstral impulse act as "shift or disparity

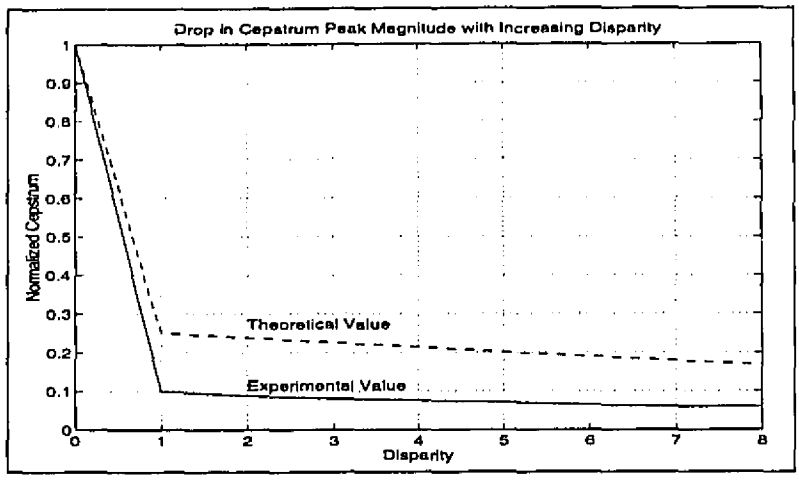


FIGURE 4.10. Decrease in the Cepstral Peak Magnitude with Increasing Disparity. The length of the original signal is 32 samples and signals are represented by concatenation.

indicators" when estimating the delay between an echo and an original. The detection of the above shift indicators for such uncorrelated signals is a trivial task, although their magnitudes decrease with increasing disparity. This is because *the only* peak or impulse present is due to disparity and easily detectable.

The salience of the autocorrelation or the cepstrum of the original signal increases with increasing self-correlation of this signal. These measures act as noise and increase the error in disparity estimation by increasing the difficulty of peak detection. Note that signal correlation causes the drop in the magnitude of the shift indicator to be less than the values given in Equations 4.13, 4.14, 4.18, and 4.19. However, the *relative* increase in the "noise" magnitude is greater than the increase in the autocorrelation peak magnitude. Therefore, the overall detection signal-to-noise ratio, represented by the relative magnitudes of the shift indicator and detection noise, is lower for correlated signals.

Further increase in self-correlation results in a signal which contains a nearly identical copy of itself. Equivalently, for such a correlated signal, *the original signal itself contains an echo.* The autocorrelation or the cepstrum of the combination of such a signal and its shifted version, contains multiple peaks or impulses representing the shifts *within each individual signal* and *between the two signals.* Using all the indicators requires a priori knowledge of their presence and number, which is impossible for arbitrary signals. Using the largest indicator may correspond to either the shift within each original signal or that between the two. The specific outcome depends on the relative magnitudes of the shifts and the

degree of inherent correlation and cannot be determined a priori. Using the representation of Equation 4.3, a signal with internal correlation can be represented as

$$\begin{aligned} g(x) &= s(x) + s(x - D_{true}) + s(x - D_{false}) \\ &= s(x) * (\delta(x) + \delta(x - D_{true}) + \delta(x - D_{false})) \end{aligned} \tag{4.20}$$

This gives rise to multiple disparity indicators in the autocorrelation or cepstrum, since multiple delayed versions exist. In the extreme, a periodic signal contains multiple copies of the same information with a constant shift between subsequent copies. This makes the distinction between "internal" and "external" echoes impossible. From a spectral point of view, periodic signals have narrow-band spectra. As explained in section 2.2.1, autocorrelation performs best for signals with a white spectrum. The performance of the cepstrum, despite its spectral whitening effect, also deteriorates with decreasing bandwidth.

Assume an arbitrary signal with an unknown degree of inherent correlation. For such a signal, autocorrelation or the cepstrum may in essence detect a disparity within the original signal, instead of one between this signal and its shifted version. Equivalently, autocorrelation and the cepstrum are susceptible to detecting *false disparities* when the signal is inherently correlated. Nonetheless, this "lack of performance" is not a shortcoming of the tool, but a direct consequence of the definition of disparity. By definition, disparity is a result of the existence of an original signal and its shifted copy. A resultant signal which contains *multiple* copies of an original produces multiple disparities. A disparity estimation method must choose one of the existing ones. The distinction between correct and false disparities becomes ambiguous for all choices based on local information. The influence of such false echo detection in both a small section and larger regions of the image is discussed later in the thesis.

6. Disparity Estimation by Addition Versus Concatenation

As discussed in Sections 2.1 and 3, the presence of an echo is indicated by the repetition of the autocorrelation function of the original signal or the presence of impulses in the cepstrum. Also recall that in addition to these "shift indicators", autocorrelation and the cepstrum contain the autocorrelation and the cepstrum of the original signal, respectively. These act as unwanted components or noise when analyzing the data. The ease with which the shift indicator is distinguished from the noise is influenced by the scheme used for combining the original and echo signals. The possible methods for combining the original

and echo signals obtained from different sensors, are addition and concatenation as explained in Section 4.

When the two signals are added, the disparity indicator - the second autocorrelation peak or the cepstral impulse - is located at a shift equal to that of the echo. For small disparities, the shift generating the echo is smaller than the width of the autocorrelation or the cepstrum of the original signal unless the signal is completely uncorrelated. Therefore, in signal combination using addition, the shift indicator for small disparities falls within the same region occupied by the autocorrelation or the cepstrum of the original signal. The latter acts as noise, making the detection of the disparity indicator difficult. Also, any signal is always correlated with itself at a shift of zero. Thus detection of non-zero disparities becomes inherently ambiguous when the original and echo signals are combined by addition.

When the two signals are placed next to each other, their disparity is increased by an amount equal to the length of each signal. If the original and echo signals are each W samples long, for disparities between $-\frac{W}{2}$ and $\frac{W}{2}$, the resultant disparity is within the range of $\frac{W}{2}$ and $\frac{3W}{2}$, when the original and echo signals are concatenated . In this range, the unwanted and decreasing autocorrelation or cepstrum of the original signal is smaller than the $-\frac{W}{2}$ to $\frac{W}{2}$ range. Therefore, for disparities smaller than one half of the signal length, concatenation causes the shift indicator to be buried in less noise than representation by addition. The disadvantage of signal representation by concatenation is the larger decrease in the peak magnitude compared to combination by addition.

We note that separate availability of the original and echo signals provides a second option for avoiding the difficulties caused by the autocorrelation or the cepstrum of the original signal in detecting the echo shift. As discussed above, it is *the autocorrelation or the cepstrum of the original signal* which acts as noise when searching for the shift indicator. One can eliminate this noise by computing the autocorrelation or the cepstrum of the original signal and subtracting the value from the resultant signal. However, due to the non-corresponding parts of the original and echo signals, the noise is slightly different from the autocorrelation or the cepstrum of each individual signal. This is seen in Figures 4.4 (g) and 4.5 (g). As seen in these figures, even after subtracting the cepstrum of the original signal, the cepstrum of the resultant contains impulses as well as unwanted components.

Using the autocorrelation or the cepstrum of the original signal may provide a solution to the problem of noise detection. However, the ambiguity in estimating non-zero disparities remains a problem for signal combination using addition. This ambiguity becomes even

more serious if small disparities are likely to occur regularly, leaving concatenation the more appropriate scheme for combining the two signals. As mentioned earlier, concatenation also eliminates the problem of noise detection for disparities less than one half of the signal length.

CHAPTER 5

The Disparity Estimation Algorithm

Chapter 4 introduced autocorrelation and the cepstrum as tools for estimating the shift generating one signal from another. Regions of a stereo image pair which contain a single disparity constitute such signals. Therefore, the stereo pair can be processed using correlation or the cepstrum to obtain its disparity profile. The cepstrum is more advantageous than correlation because it contains a distinct component which represents the shift that generates the echo. This approach is used as a cepstral filtering algorithm in [**125**]. The algorithm concatenates "patches" or "ocular dominance columns" from identical locations in the image pair. Then it computes the cepstrum of the resulting "window" and searches for its maximum value. The location of the cepstral peak indicates the shift which generates one patch from the other. Throughout this chapter the terms patch, ocular dominance column, and column refer to the region taken from each image of the stereo pair for disparity estimation. The terms window and resultant window denote the combined image region obtained by concatenating the patches from both images.

This chapter first considers the inherent assumptions and constraints of such an approach to disparity estimation. It then uses the latter, along with the properties of the cepstrum, to suggest improvements to the algorithm. The chapter is founded on the fact that disparity estimation within a single ocular dominance column is a local computation. Therefore, it only depends on the processing tool and local image properties. Later in the thesis, some of the assumptions stated in this chapter, as well as the influence of the disparities of neighbouring columns, are used to construct and interpret a disparity profile or *disparity map* of the stereo pair.

1. Underlying Assumptions of the Algorithm

As mentioned in Chapter 2, most stereo algorithms make assumptions about the compatibility of features, the uniqueness of matches, and the smoothness of surfaces in computing disparities. They also impose constraints on both the maximum detectable disparity and the viewing geometry.

It may seem that by assigning a single disparity to each image patch, the cepstral disparity estimation algorithm assumes that surfaces consist of small planes, all parallel to the viewer - not a very realistic description of the world. However, the assignment of a single disparity to a region is similar to the smoothness assumption of other algorithms. By definition of smoothness, one can always find regions small enough such that the disparity change within them remains below a specified upper bound. This gives rise to the issue of choosing the dimensions of the ocular dominance columns, a subject to be discussed later in this chapter. The fact that the disparities of neighbouring patches need not be similar is equivalent to *piecewise* smoothness. Such independence permits the occurrence of discontinuities at patch boundaries.

The cepstral disparity estimation algorithm also includes an inherent assumption about the range of existing disparities. The algorithm presupposes that the maximum disparity is a fraction (half in [**125**]) of the width of the image patch. This assumption is equivalent to that made by other stereo algorithms which define a restricted search region. Uniqueness of correspondence is also assumed in cepstral analysis by comparing every patch in one image with only one patch in the other image. If the disparities of two neighbouring columns are different, at the boundary between them, two points in one image may correspond to the same point in the other. This multiple match phenomenon is a consequence of occlusion due to depth change and occurs if one assigns a single disparity to all image points. Uniqueness exists except at those patch boundaries where a change in disparity occurs.

Disparity estimation using the cepstrum does not require perfect compatibility between the corresponding features of the images of the stereo pair. Since the cepstrum provides an overall indication of the degree of correspondence, the compatibility assumption is relaxed and the correlated regions can be slightly different. On the other hand, the usual direct correspondence determination, regardless of the process, produces a binary result. At the termination of the process, two features either correspond to one another or do not. Consequently, non-robust features that are corrupted by noise or by the difference in viewing direction, are likely not be chosen as corresponding. Conversely, the cepstrum provides an indication of similarity and is more robust to degradation or noise [**66**].

The algorithm, like most other stereo algorithms, requires opaque surfaces since specular or transparent surfaces significantly increase the complexity of the problem [54]. Finally, similar to all stereo algorithms, cepstral filtering requires and assumes a certain amount of surface markings. An image without any intensity variation is a constant (DC) signal whose cepstrum is also constant. For such a signal, of course, the concept of identical and shifted copy is ambiguous, and an infinite number of possible matches exist.

2. The Dimensions of the Ocular Dominance Columns

This section considers the factors which influence the dimensions of the ocular dominance columns, and provides a possible strategy for choosing appropriate dimensions.

2.1. Choosing the Appropriate Column Size.

2.1.1. *Approximation of the Cepstrum.* The Fourier transform and the cepstrum of the window consisting of patches from the left and right images must be *approximated* from image samples. Each ocular dominance column has to be large enough and contain enough samples so that the approximated cepstrum is a valid representation of the true value.

2.1.2. *Maximum Detectable Disparity.* Binocular disparity is estimated by measuring the location of the peak of the cepstrum of the window consisting of patches from the image pair. The size of the region in which the cepstral peak is located is of course determined by the dimensions of this composite window. Therefore, the largest disparity detectable by the algorithm is proportional to the size of the columns taken from the left and right images. The patch dimensions should be large enough to accommodate the desirable maximum detectable disparity. Note that the latter determines the column size in the units of visual angle. This is different from the cepstral approximation requirement of Section 2.1.1 which deals with column size in units of image samples (pixels).

2.1.3. *Avoiding False Disparities.* An underlying assumption behind cepstral analysis of stereo images is that an image patch and its shifted version have maximum correlation at a delay equal to the echo shift. As mentioned in Chapter 4, the image patch may contain internal correlation. Consequently, the concatenated image window may be correlated at the true disparity as well as at another shift. In the presence of such a situation, another delay, rather than the actual shift, may be chosen as the echo shift and a false disparity may be detected. For non-periodic patterns, the larger the columns, the less likely it is that *all* of their pixels are correlated at a shift other than the actual disparity. Therefore, the

columns need to be large enough and contain enough information to avoid detection of false disparities.

2.1.4. *Resolution of the Disparity Map.* The resolution of the disparity map is determined by the size of the ocular dominance columns. The resulting disparity map is coarser for larger columns. Therefore image columns have to be chosen small enough to provide a fine depth map.

2.1.5. *Avoiding Multiple Disparities.* Since a *single* disparity value is associated with each column, it is important to select columns that contain one disparity only. In other words, each column should subtend a visual angle within which surface depth undergoes only small variations.

2.2. Appropriate Column Size - The Contradictory Criteria. Sections 2.1.1 to 2.1.5 indicated that the choice of ocular dominance column size involves satisfying contradictory criteria. The approximation of the cepstrum, the detection of maximum desirable disparity, and the avoidance false disparity detection all require large column dimensions. In contrast, a high resolution disparity map and the avoidance of multiple disparities favour small column dimensions. A major challenge for the algorithm is to provide a balance among these contradictory constraints.

The above criteria can be divided into two categories: those depending on local image properties and those independent of them. The former includes false disparity detection and multiple disparity criteria, both of which depend on the nature of the processed information. The latter encompasses the remaining three criteria of cepstrum approximation, disparity map resolution and maximum detectable disparity.

A possible approach to choosing the dimensions of image columns is to use the three scene-independent criteria and set the column size a priori. The performance of the algorithm is then strongly dependent on the properties of the processed image pair. Using the features of the cepstrum, the processing tool, one can improve the performance and disassociate it from image properties as much as possible. For example, one can choose a column size based on the maximum detectable disparity and disparity map resolution criteria. With this starting point, the column size is chosen *in the units of visual angle.* To provide a proper approximation of the cepstrum, given the size of the column in visual angle units, the image resolution should be high so that each (now fixed-size) column contains enough pixels.

Therefore, a high resolution image with pre-determined maximum detectable disparity satisfies the three data-independent criteria determining the column size. However, high image resolution increases the computational cost of the algorithm. The balance between resolution and computational cost may lie in foveated images [**19**]. In foveated vision systems, the imaging device focuses on a centre of attention, referred to as the fovea, where "important" information is located. High accuracy is important in this region and not as important in the surrounding periphery. The foveated image has high resolution in the fovea and lower resolution in the periphery [**19**]. One can choose the dimensions of the ocular dominance columns so that the columns in the fovea correspond to a smaller region in visual angle units than those in the periphery. As mentioned in Chapter 3, this is in fact the same organizational strategy as the one in the ocular dominance structure of the visual cortex. Therefore, higher disparity accuracy and resolution is obtained in the fovea which is the underlying motivation for foveated vision systems. The overall visual system then ensures that the fovea is focused on those parts of the scene which are of interest. We also remember that human stereopsis in general, is limited to approximately the central 10° of the retina [**60**], [**77**]. Similarly, the active vision system under study is concerned with stereopsis in the central part of the retina.

The remainder of this chapter attempts to provide improvements to the disparity estimation tool and reduce its dependence on image properties. Then a numerical example for choosing the dimensions of ocular dominance columns is presented.

3. Improvements to the Estimation Algorithm

3.1. Re-Scaling the Cepstrum. As mentioned in Section 5.1, for finite length signals, the magnitude of the cepstral peak indicating the correct disparity decreases with increasing disparity. Equation 4.19 provided the relationship describing the ratio of the peak magnitude at any nonzero disparity d to that of zero disparity for uncorrelated signals.

3.1.1. *Re-Scaling the Cepstrum for Uncorrelated Signals.* It is possible to approximate the ratios appearing in Equation 4.19 and re-scale the value of the cepstrum at any shift by its inverse to compensate for the drop in peak magnitude. This would generate a cepstrum approximately unbiased with respect to shift, which does not favour any specific disparity range. Disparity estimation using such an unbiased cepstrum is as likely to choose a large false disparity instead of a small true one as it is to choose a small false disparity instead of a large true one.

The active vision system under study uses vergence control to fixate on the points of interest in the scene, bringing such points within the range of close-to-zero disparities. For such a system, it is preferred to avoid false disparities when estimating close-to-zero disparities at the cost of the additional likelihood of incurring error when detecting larger disparities. Denoting the event of an error in peak detection by e, and the occurrence of small and large disparities by S and L, respectively, we obtain:

$$\begin{aligned} P\{e\} &= P\{e \cap S\} + P\{e \cap L\} \\ &= P\{S\}P\{e|S\} + P\{L\}P\{e|L\} \end{aligned} \tag{5.1}$$

If $P\{S\} \gg P\{L\}$, making $P\{e|S\}$ smaller at the cost of making $P\{e|L\}$ larger decreases the overall probability of error. In other words, it is appropriate to compensate for the drop in the magnitude of the cepstral peak but preserve *some* of the bias towards zero disparity. In Equation 4.17, if one approximates the multiple self-convolution operation of $(\bullet)^{m*}$ by a self-multiplication operation $(\bullet)^m$, the ratio of the cepstral peak magnitude for any disparity d to that of disparity zero can be expressed as follows:

$$\frac{Peak_{ceps}(d)}{Peak_{ceps}(0)} = \begin{cases} 1 & d = 0 \\ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{\left(\frac{W-d}{2W}\right)^m}{1^m} = log\left(\frac{W-d}{2W} + 1\right) & d \neq 0 \end{cases} \tag{5.2}$$

when the original and echo signals are placed next to each other. Figure 5.1 illustrates this approximated value obtained from Equation 5.2, along with those of Figure 4.10.

To acknowledge the importance of surfaces near the point of fixation and zero disparity, the algorithm only partially compensates for the decrease in the magnitude of the cepstral peak. This is achieved by using the (inverse of) the ratios given by Equation 5.2 to re-scale the cepstrum of each window.

3.1.2. *Re-Scaling the Cepstrum and Data Correlation.* The relationships of Equations 4.19 and 5.2 are based on the assumption that the processed signals are uncorrelated. This assumption, although true for random-dot stereograms, rarely holds for real scenes. As a matter of fact, decreasing peak magnitude becomes an actual problem only when the signal contains inherent correlation which causes multiple cepstral peaks. As explained in Section 5.2, for such signals, a small false disparity may have a larger peak than a large true disparity. Therefore, re-scaling the cepstrum needs to be considered within the context of signals with internal correlation.

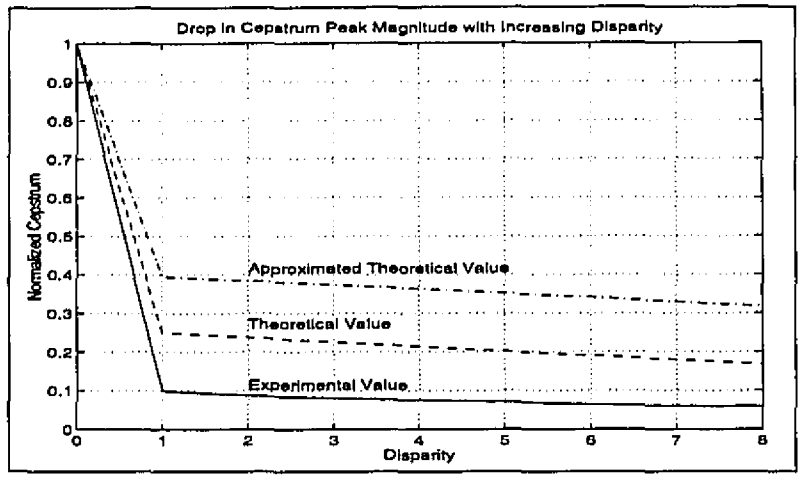


FIGURE 5.1. Approximation to the Decrease in the Cepstrum Peak Magnitude with Increasing Disparity. The length of the original signal is 32 samples and the signals are represented by concatenation. The curves illustrate the decrease in the peak magnitude as illustrated in Figure 4.10 and as described by Equation 5.2.

Using (non-negative) intensities to represent visual data results in a DC value which has a significant contribution to signal correlation. The presence of a DC value in a signal is indicated by a non-zero value at zero frequency in the power spectrum of the signal. This is re-converted to a constant value after taking the Fourier transform of the power spectrum to obtain the autocorrelation or the cepstrum. This constant value is of course, smaller in the cepstrum due to the compression property of the logarithmic transformation. Therefore, the DC component shifts the autocorrelation function and the cepstrum of a signal vertically, generating a constant level of internal correlation. Removing the DC component of every image window eliminates this contribution of signal correlation and is hence performed before computing the cepstrum.

The influence of non-DC signal correlation is dependent on the signal properties. Therefore, considering the non-DC correlation requires the introduction of the autocorrelation of the original signal as a factor into Equations 4.18 and 4.19. Attempting to estimate the image correlation properties and using them to obtain the cepstral re-scaling factor merely provides a coarse approximation to the actual values. In addition, such an approximation is computationally expensive and involves determining separate re-scaling factors for every image patch depending on the its autocorrelation function.

For signals with a decreasing autocorrelation function, the DC component is the more significant contribution to signal correlation. Furthermore, data independence of the initial

estimation stage is a primary concern of this work. Therefore, compensation for signal correlation is limited to removing the DC component of each image window.

3.2. Dimensions of Image Columns. The dimensions of the ocular dominance columns and related parameters are chosen to improve upon the algorithm cited in [**125**]. The choice of dimensions is made in units of visual angle. It is assumed that image resolution is high and each image patch contains enough samples for approximating the cepstrum.

3.2.1. *Maximum Detectable Disparity.* It is known in signal processing applications that the usefulness of correlation-type functions is limited to only a fraction of the length of the signal. This is because for larger shifts, the amount of information contributing to the value of the correlation function is so small that it is unreliable. Furthermore, as mentioned in Section 5.1, with increasing disparity the magnitude of the cepstral peak decreases to the extent that it may be buried in noise. For these reasons, the region in which the peak search is performed, or the range of detectable disparities, is reduced to $\frac{1}{4}$ of the image patch dimensions. With this decrease, maintaining the same range of detectable disparities requires doubling the dimensions of the ocular dominance columns. The width of the columns is actually doubled but the initial height is maintained. This is because, for a typical stereo vision system, the vertical disparity between the image pair is much smaller than the horizontal. Hence, halving the vertical search region does not affect the detection of the correct peak.

3.2 2. *Maintaining Disparity Map Resolution - Overlapping Ocular Dominance Columns.* Increasing the dimensions of the image columns is in contradiction to the high disparity map resolution and single disparity region requirements of Sections 2.1.4 and 2.1.5. To compensate for the loss in the resolution of the disparity map, overlapping ocular dominance columns are used.

With increasing dimensions, image patches are also more prone to including multiple or changing disparity. This shortcoming may be manifested in two ways. First, multiple disparities cause the echo power to be spread among many disparities in such a way that all peaks may be buried in noise. Second, a single disparity value is assigned to the whole ocular dominance column, which actually includes all the existing disparities. We note that the smoothness assumption makes the occurrence of either problem very unlikely. For smooth surfaces, the change in disparity within a window is small, and all of echo power is concentrated in a small section of the cepstrum. The estimated disparity is chosen in this region and reflects the true range. The second problem is further (partially) solved

by using overlapping ocular dominance columns. In practice, the same disparity is not assigned to the whole patch in the disparity map. The estimated disparity is reserved only for the central part of the ocular dominance column. The beginning and end portions of the column are, in fact, the central parts of the previous and next (overlapping) columns and receive their disparities from them. With the assumption of smooth surfaces, the "overall" or "average" disparity of the ocular dominance column is equivalent to the disparity of its central part. With such a strategy, the assigned disparity changes in harmony with the actual disparity.

Overlapping image patches, of course, increase the computational effort involved in processing each frame. In the limit, one can have image patches of width W and height H with overlaps of $W-1$ and $H-1$ in the two directions, and assign the disparity of the patch to its central pixel. In this manner, every image pixel has its own disparity value based on the shift generating its surrounding pixels in the other image of the stereo pair. However, such an approach does not exploit the surface smoothness assumption. It also results in an unnecessary increase in computational cost and no discernible benefit. The amount of overlap is chosen to provide a balance between the computational effort and the required disparity map resolution.

Avoiding the dilemma of providing such balance provides additional motivation to shy away from the unnecessary increase of the height of image columns after reducing the search region, as mentioned in Section 3.2.1. The horizontal overlap, on the other hand, is chosen as one half of the (now doubled) patch width. Such a choice maintains the initial disparity map resolution.

3.3. Removing the Cepstrum of the Original Signal. As mentioned in Section 5.2 of Chapter 4, the cepstrum of the original signal acts as "noise" when searching for the impulse which indicates the echo. Theoretically, the separate availability of the original and echo signals makes it possible to eliminate this noise. This can be achieved by computing the cepstrum of the original signal and subtracting it from the resultant signal. However, due to the non-corresponding parts of the original and echo signals, the actual detection noise is different from the cepstrum of each individual signal. Furthermore, the length of the original signal is half of that of the concatenated window. This creates a requirement for zero padding to obtain the same sampling rate in the frequency domain, as well as equal cepstral lengths. Zero padding also causes additional differences between the detection noise and the cepstrum of the original. All in all, the latter provides a mere approximation to the unwanted part of the overall cepstrum. Computing the additional cepstrum also increases

the computational cost of the algorithm significantly. Therefore, removing the cepstrum of the original signal is not performed as an improvement to disparity estimation.

3.4. Disparity Estimation with Sub-Pixel Precision. Performing a search for the largest peak in the cepstrum of the image window provides an estimation of disparity, but only to the nearest pixel. The authors in [**85**] outline a method for estimating the disparity to sub-pixel precision. They assume that the pixel-precision cepstrum is a sampled version of the sub-pixel-precision cepstrum. As a consequence, they model the sub-pixel-precision cepstrum at any point as a rectangular pulse of width one, centred around the point. The discrete cepstrum is a (sub-)sampled version of this continuous cepstrum, obtained by convolving it with a sampling function. This function is a unit rectangular pulse of width one pixel centred around the sampling point. The output of the convolution at the pixel-precision sampling point determines the value of the pixel-precision cepstrum at that point. This is schematically shown in Figure 5.2.

With such an assumption, the impulse at a true disparity, d, makes contributions to both neighbouring pixels $\lfloor d \rfloor$ and $\lceil d \rceil$. $\lfloor \bullet \rfloor$ and $\lceil \bullet \rceil$ denote the immediate smaller and larger integer numbers of the real argument. This sampling scheme simply implements linear interpolation. One can use (reverse) interpolation between the values at the two pixel-precision sampling points to obtain the location of the actual impulse. Using this method, the value of true disparity d is given by:

$$d = \lfloor d \rfloor + \frac{cepstrum\,(\lceil d \rceil)}{cepstrum\,(\lfloor d \rfloor) + cepstrum\,(\lceil d \rceil)} \tag{5.3}$$

An inherent assumption of the above approach is that the sub-pixel-precision cepstrum contains a single impulse which is the *only* source of contribution to the value of the cepstrum at the two pixel-precision sampling points. Equivalently, the approach assumes that the value of the cepstrum at each of the two pixel-precision sampling points, $\lfloor d \rfloor$ and $\lceil d \rceil$, has a contribution only from a sub-pixel-precision cepstral impulse at disparity d. With such an assumption, reverse interpolation to a single peak is justified.

The above sub-pixel estimation strategy has been adapted without any change for sub-pixel disparity estimation. The reason for choosing this simple algorithm is that the initial step in sub-pixel disparity estimation is choosing the *correct range* for the location of the peak, not the exact location. Any complex approach to sub-pixel estimation which is based on local information *after* choosing a peak is prone to the problems associated with peak detection. Likewise, any method based on the assumption that the value of the cepstrum

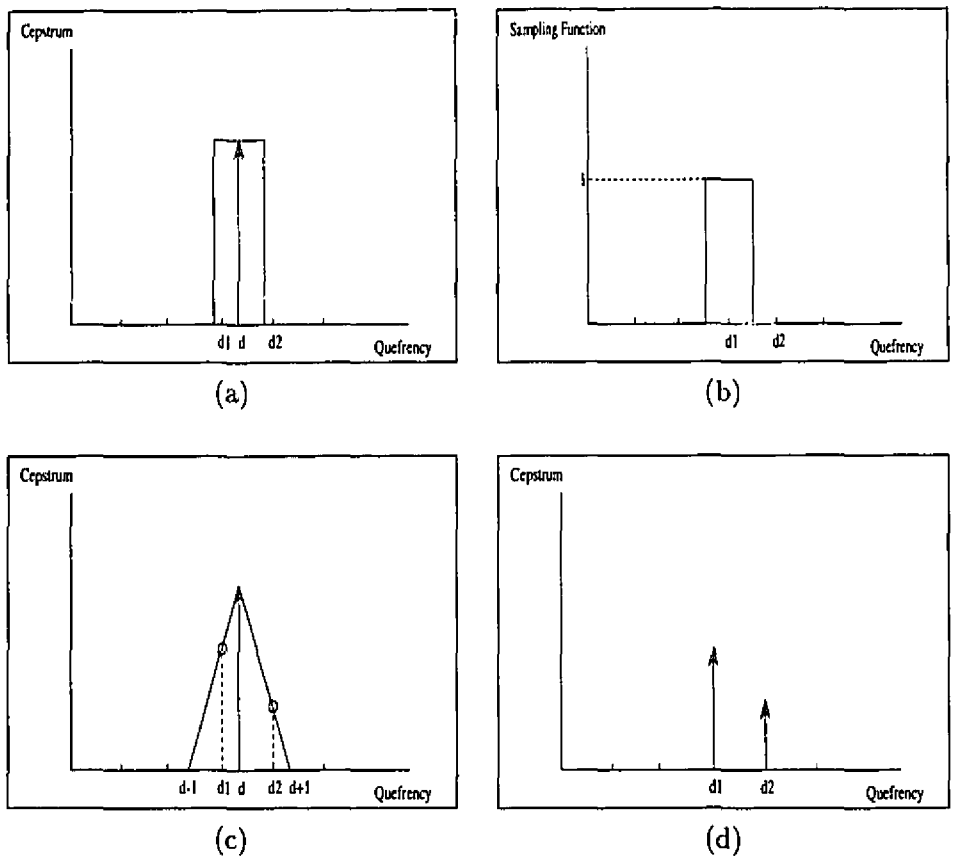


FIGURE 5.2. Disparity Estimation with Sub-Pixel Precision. (a) Assumed original sub-pixel-accuracy cepstrum. (b) Sampling function at pixel-accuracy disparity of d1. (c) Convolution of the cepstrum with the sampling functions at d1 and d2. (d) Pixel-accuracy cepstrum used in initial peak detection.

at any point is due to a single impulse is prone to the problems which arise from the distribution of power among multiple impulses. Therefore, with peak detection as the first step, increasing the complexity of sub-pixel disparity estimation offers no improvements to the uncertainty about the correct range. It may only offer little improvement to the accuracy of sub-pixel estimation.

4. An example for Choosing Ocular Dominance Column Dimensions

Ocular dominance columns with a width of 16 pixels and a height of 8 pixels provide a proper approximation of cepstrum. Assuming that a pixel spans a visual angle of $1.5'$, the width of such an ocular dominance column corresponds to a region of the retina subtending $24'$ of visual arc. Then the maximum detectable disparity of the system equals four pixels or $6'$ of the visual arc. This value is based on defining the maximum detectable disparity as

$\frac{1}{4}$ of the image patch dimensions. With a half-column overlap between ocular dominance columns, in the disparity map, each column corresponds to a visual angle of $12'$. Finally in such a system a 128-pixel $\times$ 128-pixel image covers a visual angle which is greater than $3°$. Such an image size is well within the range which is suitable for real-time processing. Also the visual angle subtended by the fovea is similar to that of many biological systems. It is interesting to note that the image resolution of this example is approximately equal to one half of the resolution of human fovea [**64**].

CHAPTER 6

From Estimated Disparities to Disparity Maps

The disparity estimation scheme of Chapter 5 provides a disparity value for each ocular dominance column. This chapter is concerned with the overall disparity map of the scene. The chapter starts by analyzing the performance of the algorithm on ocular dominance columns which contain depth discontinuities. It then studies the relationship between the disparities of neighbouring ocular dominance columns. This relationship is used for grouping many such columns into disparity regions and the construction of an overall disparity profile or map. Finally, we provide an interpretation of the overall disparity map.

1. Performance of the Algorithm at Depth Discontinuities

The smoothness assumption states that, in general, the change in disparity within small regions of an object's surface, such as those contained within an ocular dominance column, is small. Real world three-dimensional scenes are, of course, piecewise smooth and likely to include discontinuities in distance. Depth discontinuities can be divided into the two broad classes of (approximately) horizontal and non-horizontal. Once the dimensions of ocular dominance columns have been set, one can study the performance of the algorithm on image columns which contain discontinuities.

1.1. Horizontal Depth Discontinuities. At horizontal depth discontinuities, both images contain the same information about the scene. Due to occlusion, there are points in the scene that are invisible to the system. However, all such points are invisible to *both* eyes. Thus every *visible point* on either side of the discontinuity is visible to both eyes, or binocularly visible. The regions on the two sides of the depth discontinuity are, of course, marked by different disparities causing the image patch to contain more than one echo.

The sum of two different originals and corresponding echoes with shifts $D1$ and $D2$ is represented by impulses at quefrencies $D1$ and $D2$ in the cepstrum. Therefore, a window

consisting of columns from a stereo image pair with two or more disparities has a cepstrum with multiple peaks representing the disparities. The *relative* magnitudes of the peaks are determined by three factors: the relative intensities of the two sections corresponding to the two disparities, the size of the section corresponding to each disparity, and the magnitudes of the disparities themselves. As explained in Chapter 4, the magnitude of the disparity and the size of the corresponding region together determine the number of samples contributing to the cepstral peak. The expected intensity of the pixels in this region, or equivalently the surface reflectance properties, determine what the contribution of each sample is.

The cepstral filtering algorithm chooses only one of the two disparities. This is equivalent to shifting the boundary vertically towards the other. Attempting to detect the two largest peaks, of course, requires a priori knowledge of the presence of a discontinuity. Furthermore, even detecting both peaks does not give an indication of the *location* of a discontinuity. Estimating the location requires an iterative process which either reduces the height of the image column or shifts it in a vertical direction to precisely localize the discontinuity. Such an approach, although possible, is not appropriate for real-time implementation. Rather than dealing with different situations on an ad hoc basis, this thesis acknowledges their presence and addresses their consequences when interpreting the disparity map of the whole image.

1.2. Non-Horizontal Depth Discontinuities. For non-horizontal, and especially vertical discontinuities, there is a region of the scene that is visible to one eye only. Occlusion causes a region on one side of the boundary to be unmatched in one image. In a two-dimensional representation of a three-dimensional scene, the occluded region can be thought of as a stripe inserted between two regions of different disparity to compensate for the change in the shift between the two. The concept of occlusion is illustrated in Figure 6.1.

As with horizontal discontinuities, the algorithm performs normally before and after the occlusion region. For small depth changes, the width of the unmatched occlusion region is small. Therefore, there are enough points associated with one or both disparities in one image patch with matching points in the other. For such a small change the situation is analogous to the presence of a horizontal disparity and the algorithm chooses the one of the two present disparities. As a result, in the disparity map, the boundary between the two regions is shifted towards the region whose disparity is not selected.

When the change in depth is large there is a definite lack of correspondence between the ocular dominance columns from the two images. With such a lack of correspondence, the cepstrum is *not* comparing an image column to its identical and shifted version. Instead,

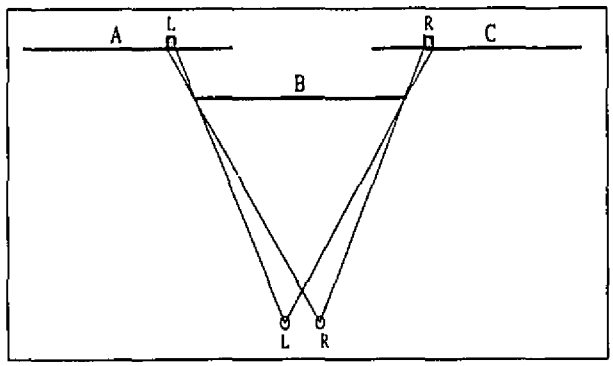


FIGURE 6.1. Occlusion. A,B, and C illustrate a cut through three surfaces at different distances. L illustrates the region of surface A visible only to the left eye. R illustrates the region of surface C visible only to the right eye. These two regions are occluded from the other eye by surface B.

the image column is being compared to a different signal. The correlation between the information in the two image columns can achieve a maximum at some unknown shift which has no physical meaning in terms of stereo correspondence. The output of the algorithm is merely a function of the statistical properties of the two image patches rather than the disparity between them. This problem, of course, can occur with any algorithm that tries to find a match for an occluded region. Once again, the influence of this issue on the overall disparity map will be discussed in the section which deals with the formation of the disparity map.

2. Thresholding the Cepstrum

If the disparity of an image patch is larger than the maximum detectable disparity, the cepstral peak falls outside of the region where peak detection is performed. With such a larger-than-detectable disparity, the value of the cepstrum within the whole peak detection range is likely to be very small. Also the distribution of the cepstrum power among multiple peaks may result in a cepstrum whose maximum magnitude is relatively small, even in the order of the detection noise. As explained above, another phenomenon which may produce such a "shallow" cepstrum and the detection of a meaningless disparity is the lack of correspondence occurring in the presence of occlusion. In all such situations, peak detection will still choose the shift with the largest cepstrum value. This is despite the fact that this largest *relative* value contains too little power to indicate the presence of an echo. Therefore, peak detection in a shallow cepstrum may result in estimating a disparity which has no physical meaning. The detected disparity is merely a function of the

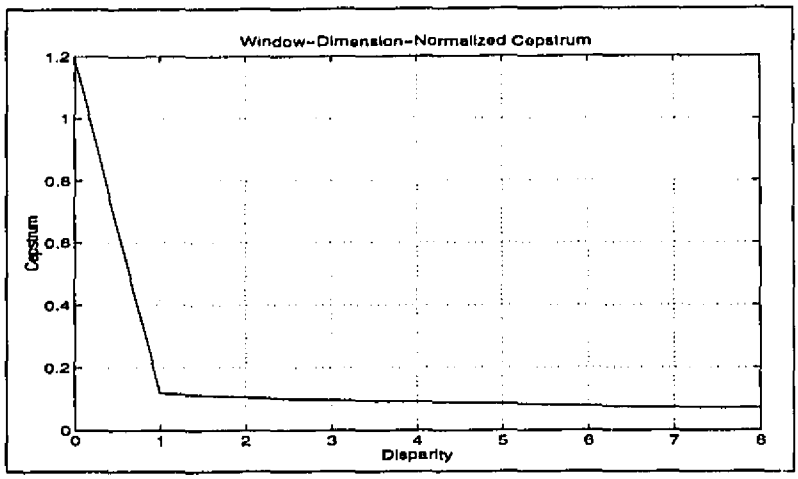
joint spectral properties of the two image patches. To avoid such situations, the computed cepstrum is thresholded before peak detection. If the magnitude of the re-scaled cepstrum is less than the threshold, no disparity information is available for that image patch.

The magnitude of the cepstrum is a function of the size of the image patch as well as the intensities of the image samples contained in the patch. Image intensities determine the power content of each sample while patch size determines the number of samples that contribute to the total cepstrum power. It is possible to approximate the expected power content of the pixels of each image patch. However, to avoid image dependent parameters, only patch dimensions are used to determine the threshold.

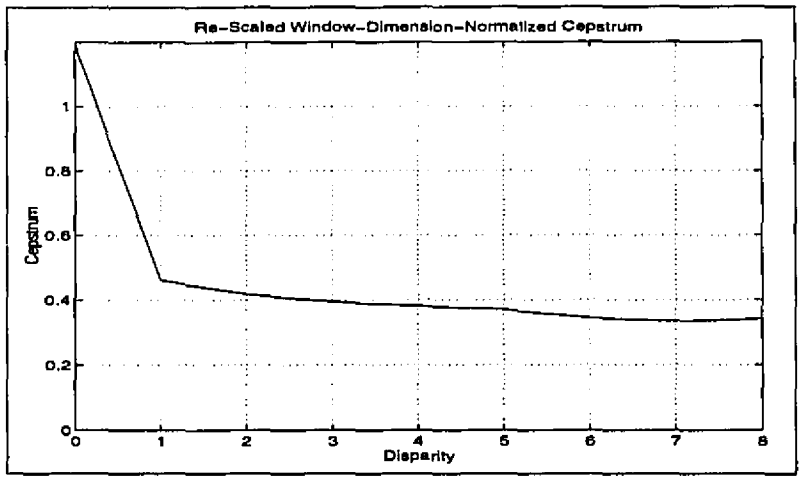
Figures 6.2 (a) and (b) illustrate the values of the cepstral peak of a zero-mean white Gaussian signal as disparity changes. The values are normalized with respect to $W \times H \times \log(W \times H)$ before and after rescaling. W denotes the width of the concatenated processing window and H is its height. The normalization factor accounts for the length of each Fourier-transformed sequence as well as the logarithmic operation involved between the two sets of transforms. This curve represents a lower bound on the cepstrum of all signals whose values have the same distribution as the signal in the figure and whose mean is non-negative. This is assuming that all of the cepstrum power is accumulated at the shift representing the disparity. Increasing the DC value shifts the whole curve upwards; any other form of internal correlation flattens the curve. In other words, assuming a Gaussian distribution, the cepstrum of any signal with the same variance as the one in Figure 6.2 lies above the curve in the figure.

The cepstral threshold is chosen as 1% to 2% of the value $W \times H \times \log(W \times H)$. For the signal of Figure 6.2, this is less than the smallest peak magnitude of the (re-scaled) cepstrum. Since the signal in this figure is a random signal, all of the cepstral power is concentrated at one peak. For other classes of signals or images, image properties such as internal correlation may result in the distribution of the cepstral power in a larger region. Choosing a threshold which is smaller than the minimum peak magnitude of the random signal accounts for the distribution of the cepstrum power. Such a distribution may result in a peak which is smaller than those in the figure but still represents a disparity. Therefore the threshold has to be small enough to accept all such peaks. A threshold less than the minimum value allows for slight decreases in peak magnitude without rejecting all the meaningful disparities (albeit at the cost of accepting some meaningless disparities).

As mentioned above, the average power content of each image sample influences the curve. However, one can assume that for any general class of applications there would most



(a)



(b)

FIGURE 6.2. Thresholding the cepstrum. The two graphs illustrate the values of the cepstral peak as disparity changes. The values are normalized with respect to the dimensions of the image window. (a) The unscaled cepstrum. (b) The re-scaled cepstrum.

likely be a specific range of image brightnesses. One could modify the cepstral threshold for different applications to account for the new scene properties.

3. Forming Disparity Regions Using Neighbouring Disparity Information

Ideally, the disparity map of a smooth surface consists of ocular dominance columns whose disparities vary slowly across the surface. Equivalently, the disparity of neighbouring ocular dominance columns of a smooth region are expected to be within a specified range of one another.

As mentioned in Chapter 4, columns which contain internal correlation may result in the detection of false disparities. Also, Section 1.2 explained that in the presence of vertical discontinuities and the associated occlusion, peak detection may result in a disparity with no physical meaning. As explained in Section 2, larger-than-maximum-detectable disparities may also result in a shallow cepstrum and a nonsense disparity. The section further illustrated that by thresholding the cepstrum before peak detection, one may be able to avoid some of these meaningless disparities. Instead they are replaced by an indication of no information. Finally, although not as significant as the previous issues, the disparity estimation noise may cause the sub-pixel accuracy disparity to be slightly different from the actual value.

Therefore, regions in the raw disparity map whose disparity is within the detectable range are corrupted by the presence of image patches containing false disparities or no information. On the other hand, regions whose disparity is outside of the detectable range - and should ideally be marked with no information - are likely to contain some patches to which a disparity is assigned.

Before using the disparity map for the symbolic representation of the scene, it is desirable to eliminate these meaningless disparities. It would also be beneficial to compensate for the effects of estimation noise on sub-pixel disparity. Patches for which no information is present require a slightly different treatment. On one hand, it is desirable to assign disparities to individual patches with no information that are embedded within patches with disparity values. On the other hand, it is appropriate to preserve large clusters of patches which contain no disparity information. The whole region can be marked as an area where disparity is larger than detectable and causes a small cepstral peak. The corresponding surfaces in the scene, are then considered to be farther or nearer than the desired range.

With the above description, we can summarize the formation of the disparity map from the output of cepstral filtering as a refining stage. The refinement involves confirming the initial estimated disparity (including no disparity), adjusting this initial value slightly, or assigning a completely new disparity value at any ocular dominance column. Furthermore, the refinement is based on the disparities of neighbouring columns. In other words, the refining stage is an "operation" in which disparity information propagates through the disparity map and the disparity of each ocular dominance column is influenced by its neighbourhood values. A possible approach to such an operation is to solve an optimization problem, in which the extremum of a cost function determines the refined disparity values. However, for

real-time applications it is desirable to avoid the high computational cost associated with optimization.

Detection of a true disparity is a result of the event in which the peak representing the true disparity is larger than that representing the false one. False disparities are of course an outcome of the inverse of the above event. It is reasonable to assume that the internal correlation which causes the detection of false disparities merely illustrates a *limited degree* of similarity to the original signal. Thus, one can infer that the likelihood of the occurrence of event that a true disparity is detected is greater than that of detection of a false disparity. Since the estimation of a false disparity is less likely than a true one, the "defects" of the disparity map occur as outliers. With such an assumption, it is possible to take a more efficient approach to the refining process. Eliminating noise which appears as outliers in the original data is suited to processing with a median filter. Besides eliminating the disperse noise, median filtering also serves to adjust those values that are slightly different from the "collective" disparity of their neighbours.

3.1. Modified Median Filtering. To account for the presence of columns with no disparity information, the traditional median filtering is slightly modified. In the modified scheme, a contribution to the output of the filter is made either by those points which contain a disparity value or those which have no disparity information, *but not both.* In other words, the value of any filtered point equals the median of *only* those neighbouring points to which a disparity is assigned unless more than one half of the neighbouring points contain no-disparity information. Under the latter condition, the filtered point is marked as a point with no disparity. This is an illustration of the (above-mentioned) notion of considering large regions with no disparity information as areas where the cepstral peak is too small to indicate a detectable disparity. "Large region" is defined to consist of more than one half of the area covered by the median filter.

3.2. Two Different Approaches to Median Filtering. The cepstral filtering algorithm assigns a single disparity value to each ocular dominance column. This value is then mapped to all the pixels which fall within the boundaries of the column, while taking overlapping columns into consideration. The mapping operation projects the *single* disparity of the ocular dominance column into *multiple* copies of this value as the disparities of the pixels in the column. With such a mapping operation, one can consider two different sets representing the disparities of the whole image: one before the mapping and the other after. The former set is that of the disparities of ocular dominance columns and the latter

set is that of disparities of individual image samples. The first set associates a single disparity with each ocular dominance column. The second set assigns as many (identical) disparities to the column as the number of pixels in it. Therefore the former set, hereafter referred to as the data domain, represents the disparities *before mapping to the pixels of individual ocular dominance columns.* It has a membership equal to the total number of ocular dominance columns in the image. The latter set, referred to as the image domain, is that of disparities *after having been mapped to the pixels of individual ocular dominance columns.* Its membership equals the number of pixels in the image.

It is possible to perform the refinement of the estimated disparity values in either the data domain or the image domain. The two schemes respectively correspond to applying the median filter *before* and *after* mapping the disparity of each ocular dominance columns to its pixels. Refinement in the data domain, considers the influence of the disparities of neighbouring ocular dominance columns and not pixels. Therefore, it assumes that all the pixels in a column are equally influenced by all the pixels in the neighbouring columns regardless of the relative positions of individual pixels in each column. Assume that the size of the median filter in units of visual angle is the same in both schemes. Then the data domain refinement is equivalent to performing the image domain refinement on the central pixel of each column and using this value for all the pixels within that column. The two methods are schematically illustrated in Figure 6.3.

Disparity refinement in the image domain has the added advantage of considering the relative position of each pixel *within the ocular dominance column* when using the neighbouring information. The neighbourhood of each sample comprises those pixels that are within a specified distance from that pixel. Considering the disparity and neighbourhood information for each individual pixel causes any disparity change to occur at the individual pixel level. On the other hand, filtering all the pixels of an ocular dominance column simultaneously causes the disparity change to occur at the column boundaries only. Therefore, the image-domain refinement of the disparity map imposes a smoothness constraint on the boundaries between disparity regions. This is a result of "spreading" any change in disparity over the whole ocular dominance column.

3.3. Dimensions of the Median Filter. The dimensions of the median filter determine the extent of the neighbourhood region which is used in refining each disparity value. Increasing the size of the neighbourhood region corresponds to the assumption that disparity information remains within a specified range in a larger area. In other words, larger filter dimensions impose the surface smoothness assumption more strongly.

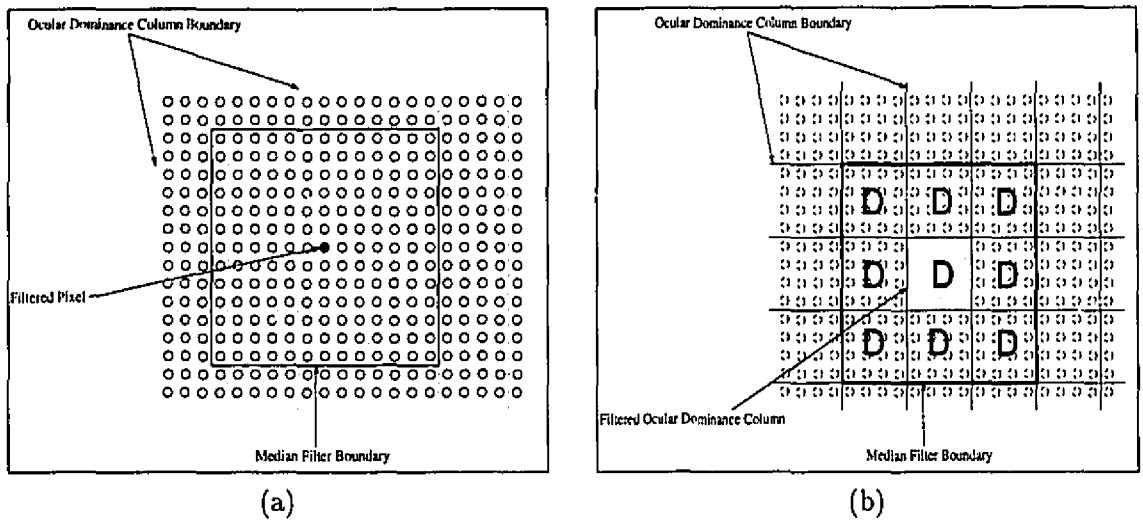


FIGURE 6.3. Disparity Refinement in the Image and Data Domains. (a) Image domain filtering. A disparity is assigned to each pixel in the ocular dominance column. Every pixel in every ocular dominance column is filtered separately. Filter boundaries need not coincide with ocular dominance boundaries. (b) Data Domain filtering. Only a single disparity is associated with each ocular dominance column. The whole column is filtered as a single entity. Filter boundaries always coincide with ocular dominance column boundaries.

The choice of filter size has to provide a balance between using neighbourhood information for disparity refinement and allowing gradual changes in surface disparities. For this reason, the chosen neighbourhood region consists of 1.5 to 2 ocular dominance columns in each direction for both image and data domain filtering. In the image domain filtering, the neighbourhood of each pixel is determined by the filter size and its relative position within the ocular dominance column. In the data domain filtering, the neighbourhood of each ocular dominance column always includes the column itself as well all the columns that are in contact with the filtered one.

4. The Refined Disparity Map and Figure-Ground Separation

The principle assumptions of the refinement stage are that the surfaces in the scene are piecewise smooth and false disparities occur as outliers. With such assumptions, the refined disparity map consists of disparity regions each of which represents a smooth surface in the scene. Since the algorithm assigns a single disparity to each column in the estimation stage, the surface boundaries may have been shifted by no more than the dimension of an ocular dominance column. Sub-pixel estimation and median filtering also illustrate the changes within a single surface. Therefore, in the overall map one can distinguish the locations of different surfaces and an approximation of their individual properties.

A symbolic representation of the scene with such characteristics, forms the basis for figure-ground separation using stereopsis. Such a description distinguishes between neighbouring surfaces based on *difference in depth between them* (and not the absolute distance of each surface). It further provides a coarse estimation of the surfaces themselves. Such output is in harmony with the figure-ground separation role defined in [**42**]. It also agrees with the view of [**116**] that complete surface reconstruction occurs only after an initial description, based on non-constant depth changes, is obtained.

CHAPTER 7

Experimental Results

This chapter is devoted to studying the method developed for obtaining the disparity map of a stereo pair within an experimental context. The chapter first examines the disparity estimation and disparity map construction stages separately. For each stage the influences of the different aspects of the algorithm are studied. Examples include the effects of the re-scaling of the cepstrum or the domain in which median filtering is performed. After studying all aspects of the method individually, the overall disparity maps obtained by applying the complete method to various image types are presented and explained.

One of the images used for obtaining the overall disparity map is a random-dot stereogram in which regions of each image are *identical* and shifted versions of ones in the other image. Experiments are also performed on synthetic images of three-dimensional scenes. Synthetic images are those which have not been obtained with actual cameras from real three-dimensional scenes. Nevertheless, such images are based on descriptions of three-dimensional objects. Also the imaging process is mimicked in their generation. Consequently, these images share all of the properties of real stereo pairs, especially the differences which exist between left and right images due to different viewing angles. Such dissimilarities include variations in lighting as well as foreshortening. Finally the output of the algorithm on real stereo images of three-dimensional scenes is studied.

Historically, various stereo algorithms have been developed based on different assumptions about the role of stereopsis and the type of images they deal with. Such assumptions often represent the specific application for which stereopsis is used. As a result, many of these algorithms have evolved to provide a particular type of result on a specific class of images. Many feature-based algorithms, for example, may perform poorly on images with a great deal of texture but outperform region-based algorithms on surfaces with little texture. Therefore the performance and the computational cost of each algorithm is dependent on

the specific context for which it has been developed. For this reason this chapter does not compare the experimental results obtained to the performance of other classes of stereo algorithms. The algorithm and its performance are studied while considering the assumptions that this thesis makes about the role of stereopsis and surfaces in the world.

1. From Stereo Image Pairs to Disparity Maps

Figure 7.1 contains three stereo image pairs used in the experiments designed to illustrate the role of each aspect of the algorithm. The depth profiles for these image pairs are schematically shown in Figure 7.2.

The image pair in Figure 7.1 (a) contains two objects: a cube and a sphere. The baseline between the two cameras is parallel to the surface of the cube and the point of fixation lies on this surface. Therefore, the surface of the cube has a disparity of approximately zero. The sphere is located farther from the cameras than the cube and its disparity is greater than zero. Due to the change in depth along the surface of the sphere there is a disparity change of approximately one pixel between the central point and the boundary of its surface. The background is the farthest surface and has the largest disparity of the three.

The stereo pair of Figure 7.1 (b) contains a stair-case depth profile consisting of surfaces at five different distances, including the background. The leftmost surface contains the point of vergence and has a disparity of zero. Disparity increases incrementally from left to right. The background consists of the whole upper half of the image as well as the far right section of the lower half.

Finally, the image pair in Figure 7.1 (c) contains a cone whose vertex points outwards, against a background of constant depth profile. Because of the linear depth profile of the cone there is a constant change in disparity along its surface. There is a sudden change in disparity between the base of the cone and the background.

1.1. Disparity Estimation Algorithm.

1.1.1. *Re-Scaling of the Cepstrum.* Figure 7.3 (a) shows the output of the algorithm described in [**125**] on the first two test images of Figure 7.1. In both pairs, the maximum disparity belongs to the background and has a value of four pixels. The width of each ocular dominance column is eight pixels, with the search region defined as one half of this value as in [**125**]. Figure 7.3 (b) illustrates the effect of re-scaling the cepstrum before peak detection.

As seen in this figure, disparity estimation without re-scaling results in detecting false disparities in ocular dominance columns with large disparities. In particular, for disparities

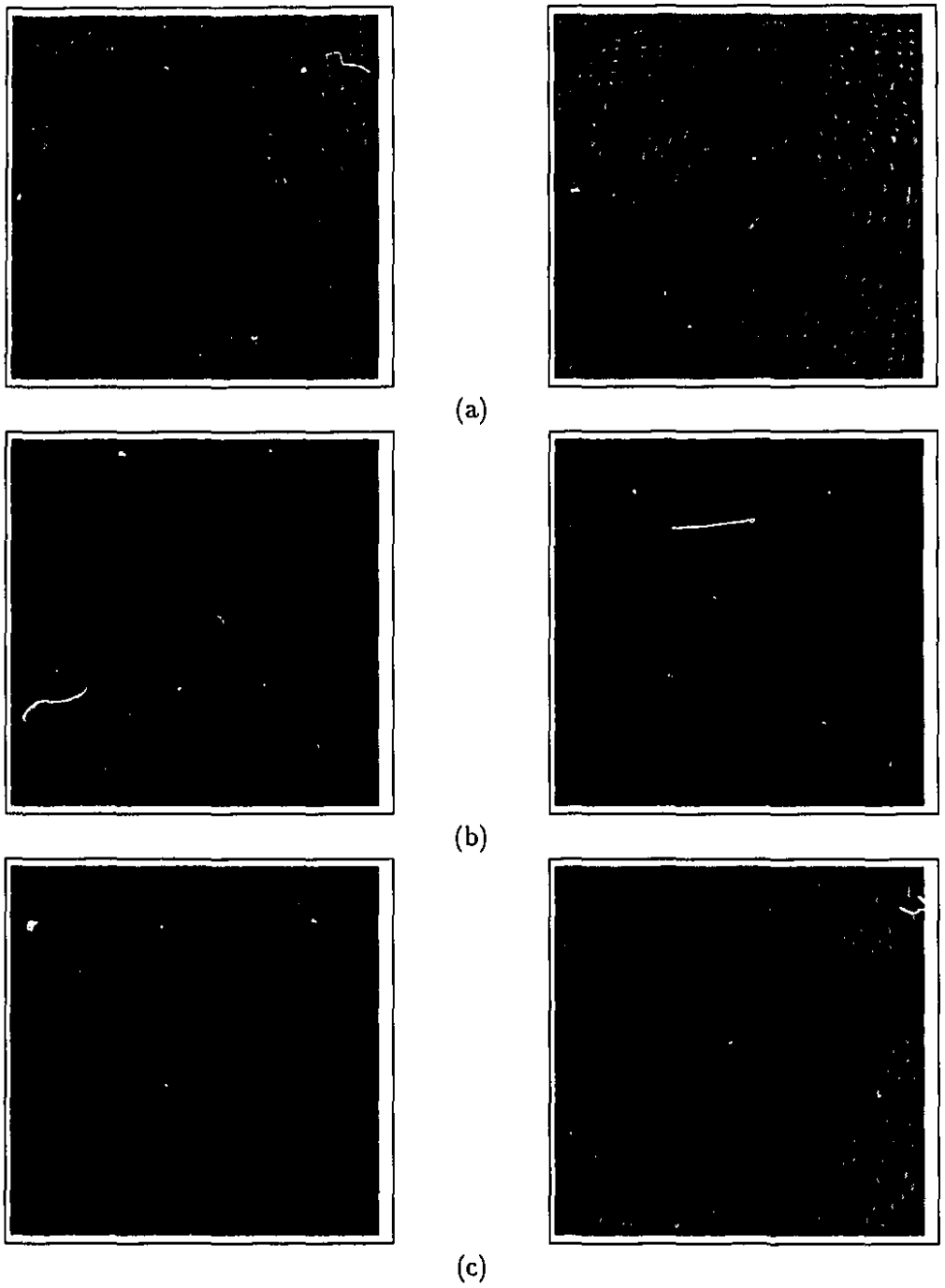
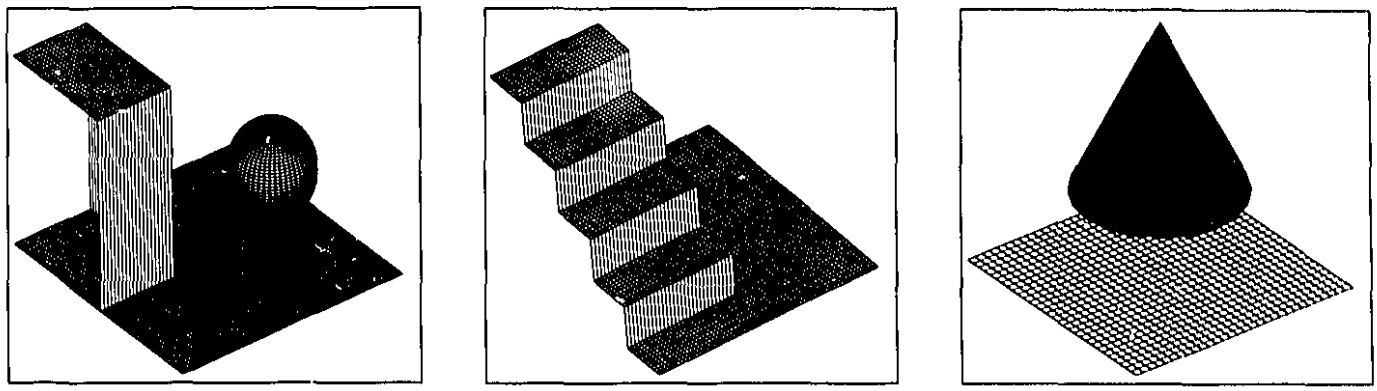


FIGURE 7.1. Stereo Image Pairs. (a) Objects. (b) Staircase. (c) Cone.



(a) (b) (c)

FIGURE 7.2. Depth Profiles for Stereo Image Pairs. The depth maps are not to scale. (a) Objects. (b) Staircase. (c) Cone.

of three and four pixels - that are greater than one quarter of the patch width - distinguishing between the actual disparity of the region and estimation noise becomes impossible. Re-scaling the cepstrum partially reduces the extent by which small disparities are favoured. This in turn reduces the level of noise in large disparity regions at the cost of introducing a few false disparities in regions where the disparity is small.

1.1.2. *Accounting for Signal Correlation Due to the DC Component.* As explained in Section 3.1.2 of Chapter 5, the mean value of each image window is removed to eliminate the contribution of the DC value to signal correlation. The effect of removing the DC value of the composite window before computing the cepstrum is shown in Figure 7.4. Subtracting the signal mean removes the DC component of signal correlation. This makes the use of re-scaling factors of Equation 5.2 which are developed for uncorrelated signals more appropriate.

1.1.3. *Maximum Detectable Disparity.* As mentioned in Chapter 5, for a given image resolution, the dimensions of the ocular dominance columns can be chosen based on the desired maximum detectable disparity. Therefore, the dimensions of the ocular dominance columns are not considered as independent variable factors, but a function of image resolution and maximum detectable disparity.

Figure 7.5 contains the initial or unfiltered disparity map after reducing the size of the region in which the search for the cepstral peak is performed from one half to one quarter of the width of the ocular dominance columns.

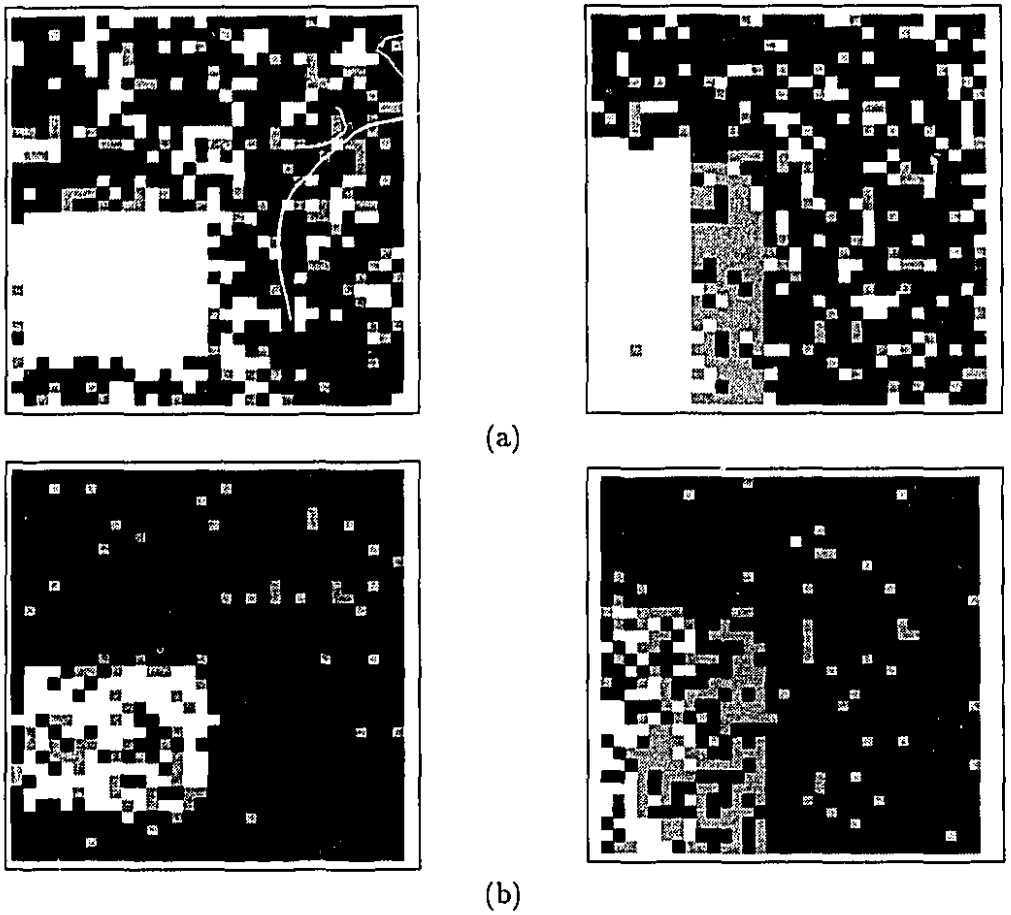


FIGURE 7.3. Re-scaling the Cepstrum and Disparity Estimation. The left hand side illustrates the disparity maps for Figure 7.1 (a) and the right hand side for Figure 7.1 (b). Peak detection is performed in a region whose width is one half of the width of the ocular dominance column. (a) Initial disparity map without re-scaling the cepstrum using the algorithm described in [125]. (b) Initial disparity map after re-scaling the cepstrum.

After halving the peak search region, maintaining the same maximum detectable disparity requires doubling the width of the ocular dominance columns. The width of the ocular dominance columns used to obtain the disparity maps of Figure 7.5 is sixteen pixels compared to eight pixels in Figure 7.4. As seen by the result, decreasing the peak detection region considerably reduces the estimation error at the cost of decreasing the resolution of the disparity map.

1.1.4. *Overlap of the Ocular Dominance Columns and the Disparity Map Resolution.* Overlapping ocular dominance columns solve the problem of reduced resolution at the cost

FIGURE 7.4. Removing the Signal Mean before Disparity Estimation. Peak detection is performed in a region whose width is one half of the width of the ocular dominance column and the cepstrum is re-scaled. (a) Disparity map for Figure 7.1 (a). (b) Disparity map for Figure 7.1 (b).

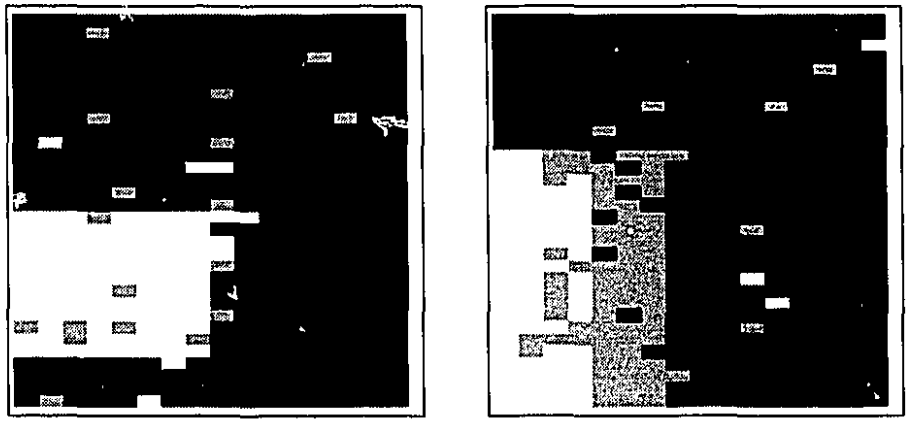


FIGURE 7.5. The Effect of Reducing the Maximum Detectable Disparity. The search region is reduced from one half to one quarter of the (doubled) ocular dominance column width. The signal mean is removed and the cepstrum is re-scaled. (a) Disparity map for Figure 7.1 (a). (b) Disparity map for Figure 7.1 (b).

of increased computational effort. The influence of increasing the (horizontal) overlap on disparity map resolution is illustrated in Figure 7.6.

Section 3.2.2 of Chapter 5 also explained that overlapping columns compensate for the presence of changing disparities in ocular dominance columns. This is done by ensuring that the estimated disparity is assigned only the central part of each patch. The effect of overlapping columns on surfaces with changing disparity is illustrated in Figure 7.7 for the image of a cone whose depth and disparity undergo a constant change.

One can see that increasing the overlap of ocular dominance columns localizes the transition between different disparities more precisely. This is of course because the estimated disparity is assigned only to the central part of the patch. With the assumption of smooth surfaces, the disparity of this central region is equivalent to the "overall" or "average" disparity of the patch.

1.1.5. *Removing the Cepstrum of the Original Signal.* The most significant cause of error in disparity estimation is the problem of false disparities due to internal correlation of the original signal. Internal correlation is manifested by extrema in the cepstrum of the

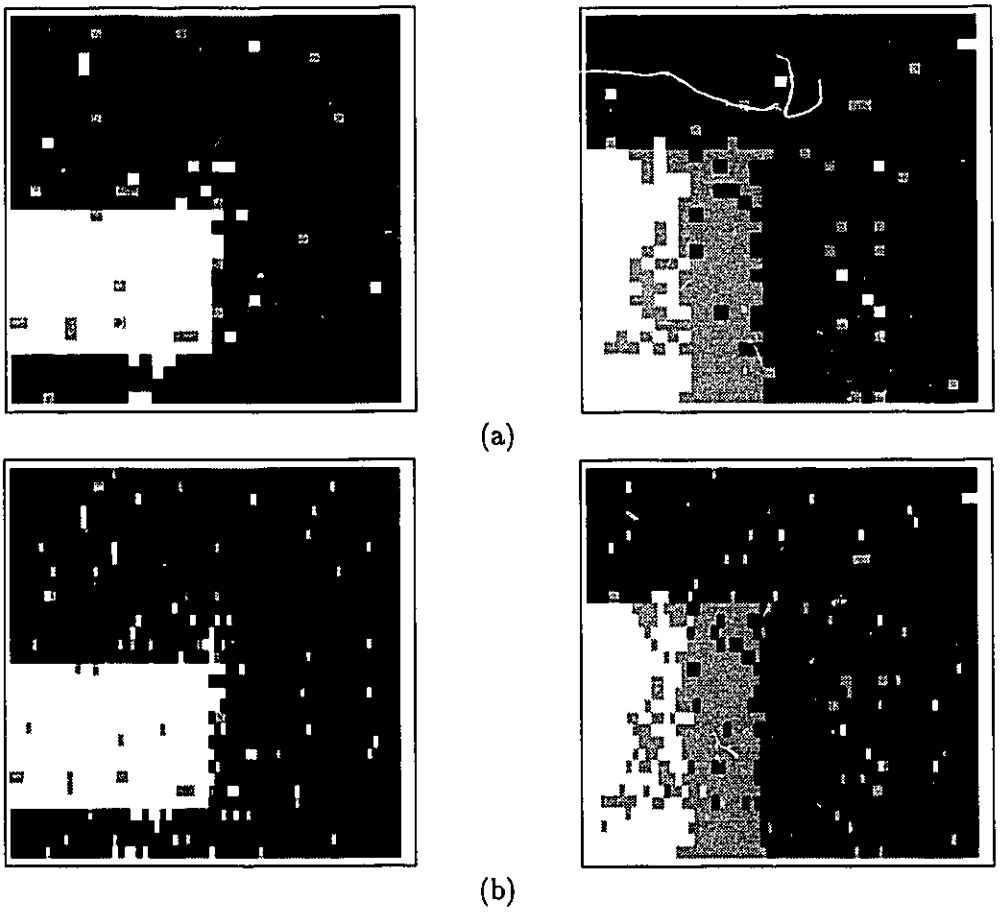


FIGURE 7.6. Ocular Dominance Column Overlap and the Resolution of the Disparity Map. The left hand side illustrates the disparity maps for Figure 7.1 (a) and the right hand side for Figure 7.1 (b). Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled. (a) Overlap equals 50% of the column width. (b) Overlap equals 80% of the column width.

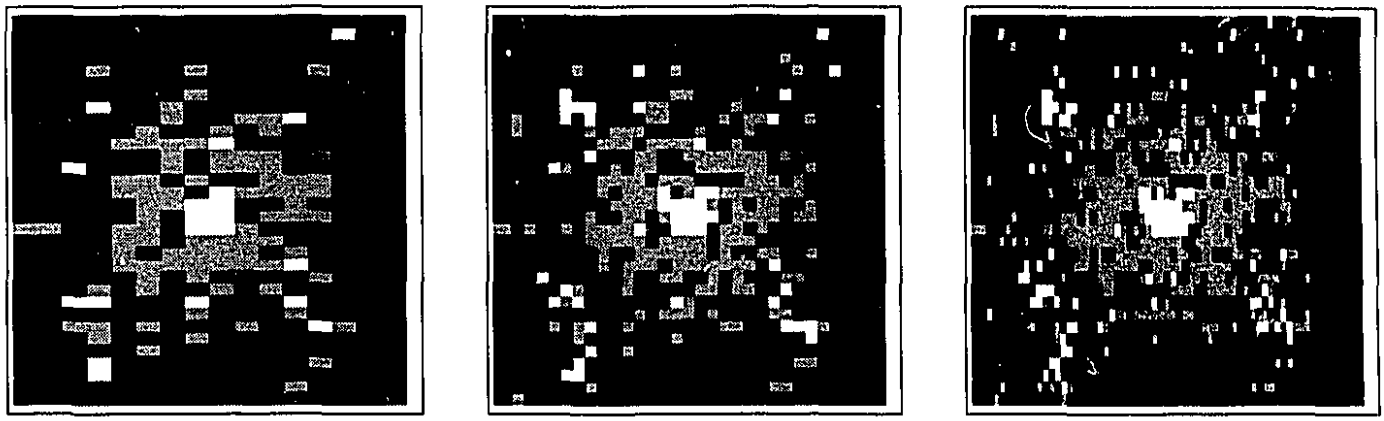


FIGURE 7.7. The Effect of Ocular Dominance Column Overlap on Preserving Disparity Gradient. All disparity maps correspond to Figure 7.1 (c). Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled. (a) No overlap. (b) Overlap equals 50% of the column width. (c) Overlap equals 80% of the column width.

original signal which act as "noise" in detecting the cepstral peak. Figure 7.8 illustrates the influence of removing the cepstrum of the original signal before peak detection.

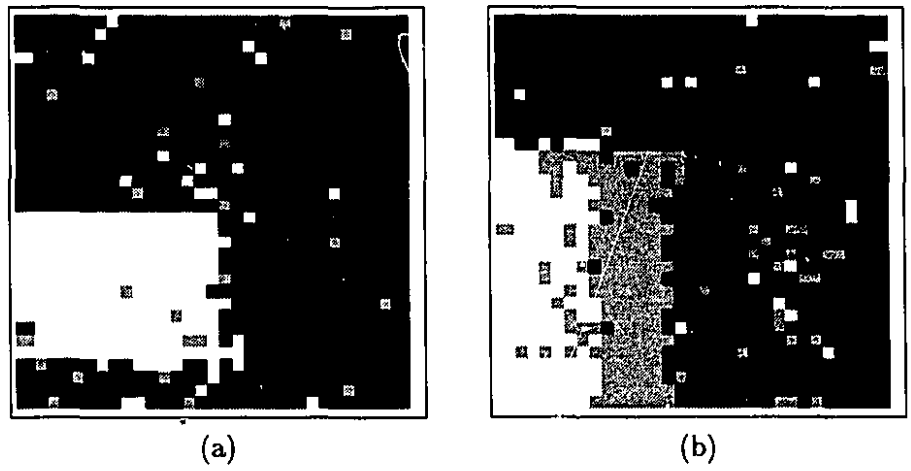


FIGURE 7.8. Removing the Cepstrum of the Original Signal. Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled. The overlap between neighbouring columns is 50%. (a) Disparity map for Figure 7.1 (a). (b) Disparity map for Figure 7.1 (b).

The result illustrates that this operation does not offer any improvement to the performance of the algorithm. As explained in Chapters 4 and 5 this is due to the difference

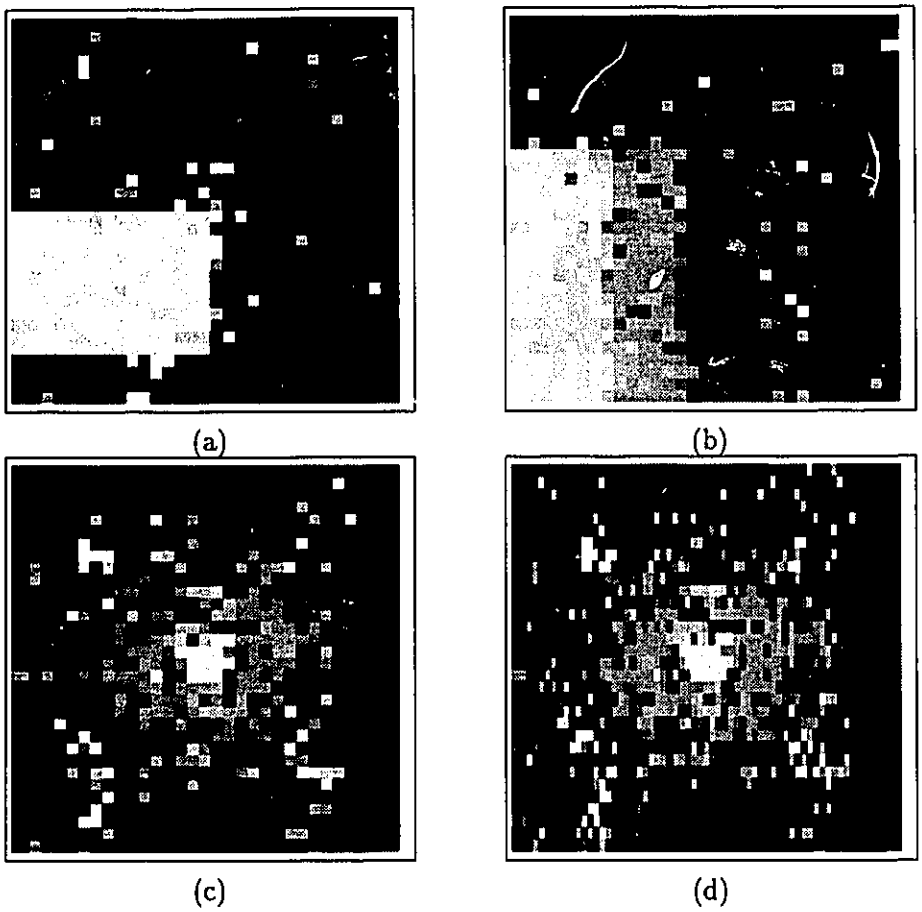


FIGURE 7.9. Disparity Estimation with Sub-Pixel Accuracy. Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled. (a) (b), and (c) The overlap between neighbouring columns is 50%. (d) The overlap between neighbouring columns is 80%. (a) Disparity map for Figure 7.1 (a). (b) Disparity map for Figure 7.1 (b). (c) and (d) Disparity maps for Figure 7.1 (c).

between the cepstrum of each of the two images and the detection noise. The difference is due to the non-corresponding parts of the two ocular dominance columns.

1.1.6. *Disparity Estimation with Sub-Pixel Accuracy.* Sub-pixel disparity estimation is performed by interpolating between the largest cepstral peak and the larger of its adjacent peaks. The results of estimating disparity with sub-pixel accuracy is illustrated in Figure 7.9. In particular, Figure 7.9 (d) shows the gradual change in the disparity of the surface of a cone and illustrates the effect of this approach in estimating sub-pixel disparities.

1.2. Disparity Map Construction.

1.2.1. *Thresholding the Cepstrum.* The result of thresholding the cepstrum before peak detection is illustrated in Figure 7.10 (a). Comparing these disparity maps with those in Figures 7.9 (a) and (b) demonstrates that thresholding the cepstrum results in replacing some of the false disparities by an indication of no disparity information. This is particularly true for those false disparities that occur at depth discontinuities.

The effect of varying the threshold level is shown in Figures 7.10 (b) and (c). A comparison between the disparity maps of Figure 7.10 (a) and those of Figures 7.10 (b) and (c) shows that a threshold of $0.02 \times W \times H \times \log(W \times H)$ flags the majority of the false disparities as unknown locations without affecting many of the true ones. W and H are the dimensions of the window resulting from concatenating the ocular dominance columns taken from the two images.

1.2.2. *Refining the Initial Disparity Map.* Median filtering is used to refine the initial disparity estimations, including the presence of no disparity information, using the disparities of the neighbours. Section 3.2 of Chapter 6 explained that filtering can be performed using either individual pixels or individual ocular dominance columns as filtering kernels. Figures 7.11 and 7.12 illustrate the results of filtering the thresholded disparity maps of the test figures in both image and data domains.

The results illustrate that false disparities occur as isolated noise for these image pairs. Median filtering serves to eliminate these false disparities in the disparity map. Furthermore, Figure 7.11 illustrates that performing the filtering operation at every pixel results in smoother transition between regions of different disparity. Finally, Figure 7.12 illustrates the smooth transition property for a surface with a changing disparity profile.

1.2.3. *The Size of the Median Filter.* Section 3.3 explained that the dimensions of the median filter correspond to the degree of the assumed surface smoothness. The influence of the filter size is illustrated in Figure 7.13.

As seen in the figure, increasing the size of the median filter increases the neighbourhood support as well as the smoothness of the boundaries between the disparity regions.

2. Performance of the Overall Method

The previous two sections studied the influences of the various aspects of the method experimentally. The results of the experiments illustrate the effectiveness of the improvements to the algorithm proposed in Chapter 5, as well as the performance of median filtering

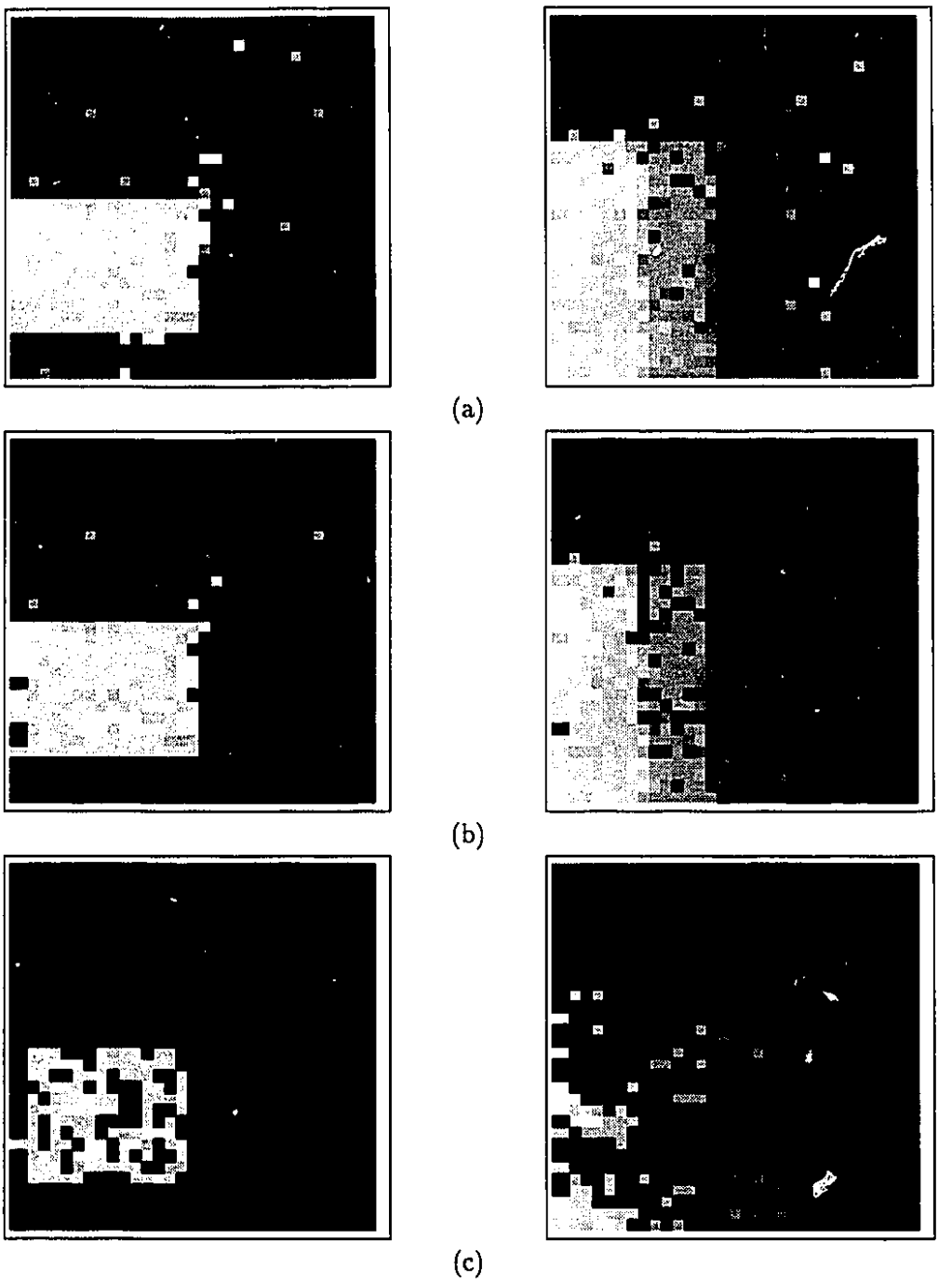


FIGURE 7.10. Thresholding the Cepstrum Before Peak Detection. The left hand side illustrates the disparity maps for Figure 7.1 (a) and the right hand side for Figure 7.1 (b). Black represents the presence of no disparity information. Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal is removed and the cepstrum is re-scaled. The overlap between neighbouring columns is 50%. The cepstrum of the original signal is removed. (a) Threshold is $0.02 \times W \times H \times \log(W \times H)$. (b) Threshold is $0.04 \times W \times H \times \log(W \times H)$. (c) Threshold is $0.08 \times W \times H \times \log(W \times H)$.

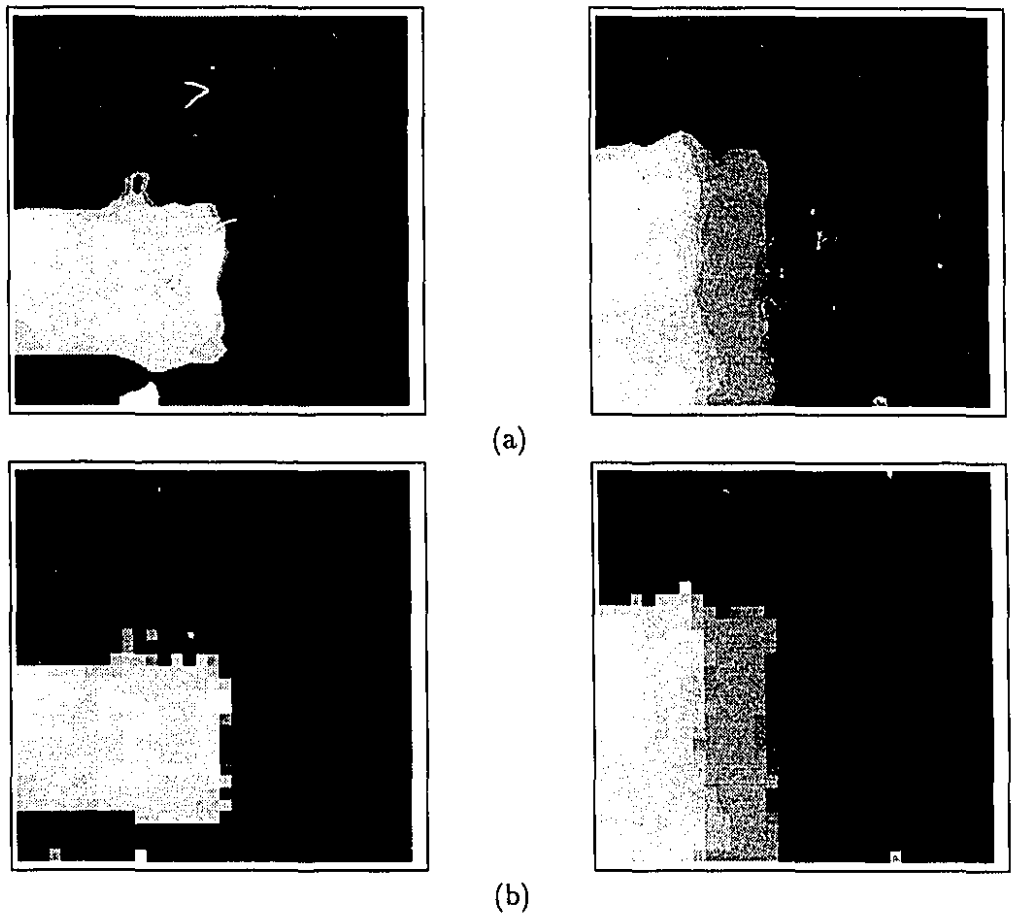
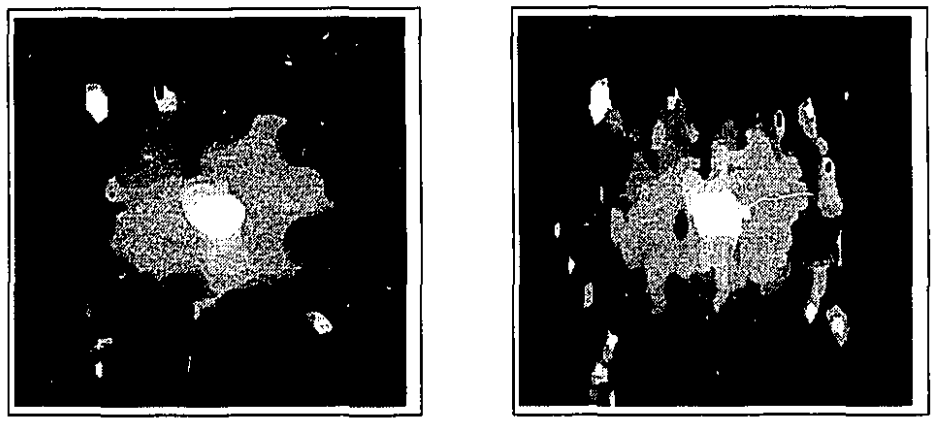


FIGURE 7.11. Refinement of the Disparity Map Using Median Filtering. The left hand side illustrates the disparity maps for Figure 7.1 (a) and the right hand side for Figure 7.1 (b). Black represents the presence of no disparity information. Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled and thresholded. (a) Filtering in the image domain. The dimensions of the filter are three times those of the ocular dominance columns (after overlap). (b) Filtering in the data domain. The width of the filter is three.

in refining the initial disparity map. This section examines the attainment of the complete scheme on different stereo image pairs. The results of these experiments, along with those in Figures 7.11 and 7.12, provide an insight into the method which can be used as a guideline for its application as a part of a vision system.

2.1. Performance on Random-Dot Stereogram. Figures 7.14 (a) and (b) show the left and right images of the random-dot stereogram used in the experiment. The stereo signal contained in the image pair is shown in Figure 7.14 (c). The stereo signal represents



(a) (b)

FIGURE 7.12. Median Filtering and Changing Disparity. Disparity maps correspond to Figure 7.1 (c). Black represents the presence of no disparity information. Peak detection is performed in a region whose width is one quarter of the width of the ocular dominance column. The signal mean is removed and the cepstrum is re-scaled and thresholded. (a) The overlap between neighbouring columns is 50%. (b) The overlap between neighbouring columns is 80%.

a disparity of 10 in the central region of the image pair. Figure 7.15 contains the computed disparity map for the random-dot stereogram of Figure 7.14.

Comparing the result with the actual disparity map of Figure 7.14 (c) illustrates that the algorithm has correctly obtained the disparity map of the pair. In the initial disparity map some of the ocular dominance columns located on vertical depth discontinuities contain false disparities. This is due to the lack of corresponding information in the columns from the left and right images caused by occlusion. The boundaries between the two disparity regions are slightly shifted due to the fact that a single disparity is assigned to the whole region covered by an ocular dominance column. The overall disparity map clearly represents the boundaries between surfaces of different depths and is suitable for figure-ground separation based on depth.

2.2. Performance on Real Stereo Image Pairs. Some of the stereo image pairs for the experiments have been obtained from an available stereo image bank compiled by SRI International. Many such images have been taken for use with specific algorithms. For example, most traditional stereo algorithms search for corresponding features in the two images of the stereo pair. To simplify the search, such algorithms use aligned cameras which result in horizontal disparities and therefore horizontal epipolar lines only. Two imaging devices are aligned if they have identical vertical elevation and orientation. Another consequence of camera alignment is the rare occurrence of zero and small disparities. By

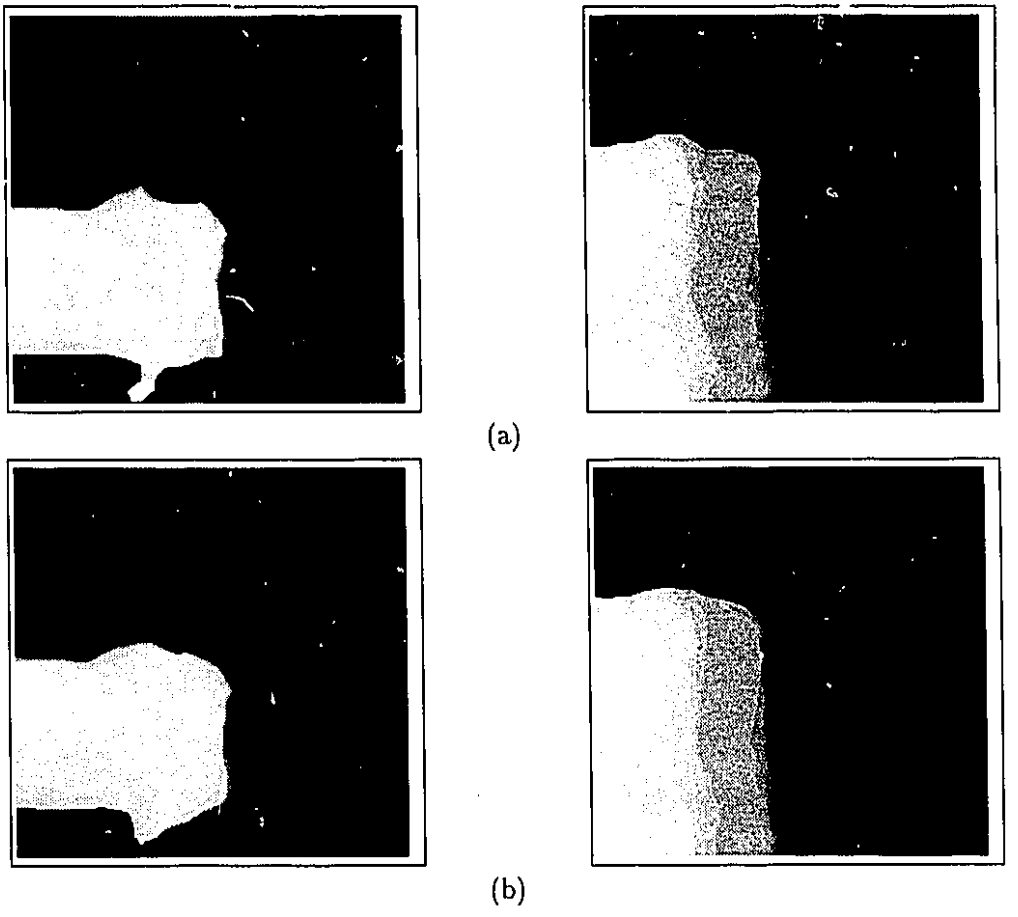


FIGURE 7.13. Median Filter Dimensions. The left hand side illustrates the disparity maps for Figure 7.1 (a) and the right hand side for Figure 7.1 (b). (a) The dimensions of the filter are four times those of the ocular dominance columns (after overlap). (b) The dimensions of the filter are five times those of the ocular dominance columns (after overlap). The search region is one quarter of the ocular dominance column width. The signal mean is removed. The cepstrum is re-scaled and thresholded. The overlap between neighbouring columns is 50%. The cepstrum of the original signal is removed.

using the two-dimensional cepstrum, the algorithm described in this thesis can easily handle vertical disparities and hence is not restricted to parallel-axis imaging devices. By removing the alignment restriction and the use of vergence, the imaging system can fixate on the regions of interest and reduce their disparities to the range detectable by the system. This is similar to the saccading and fixation movements of the human visual system. However, in the described experiments the range of detectable disparities has been based on the available aligned stereo image pairs.

Figures 7.16 to 7.18 contain stereo image pairs and the initial and refined disparity maps obtained by segmenting the pair using the cepstral disparity estimation algorithm

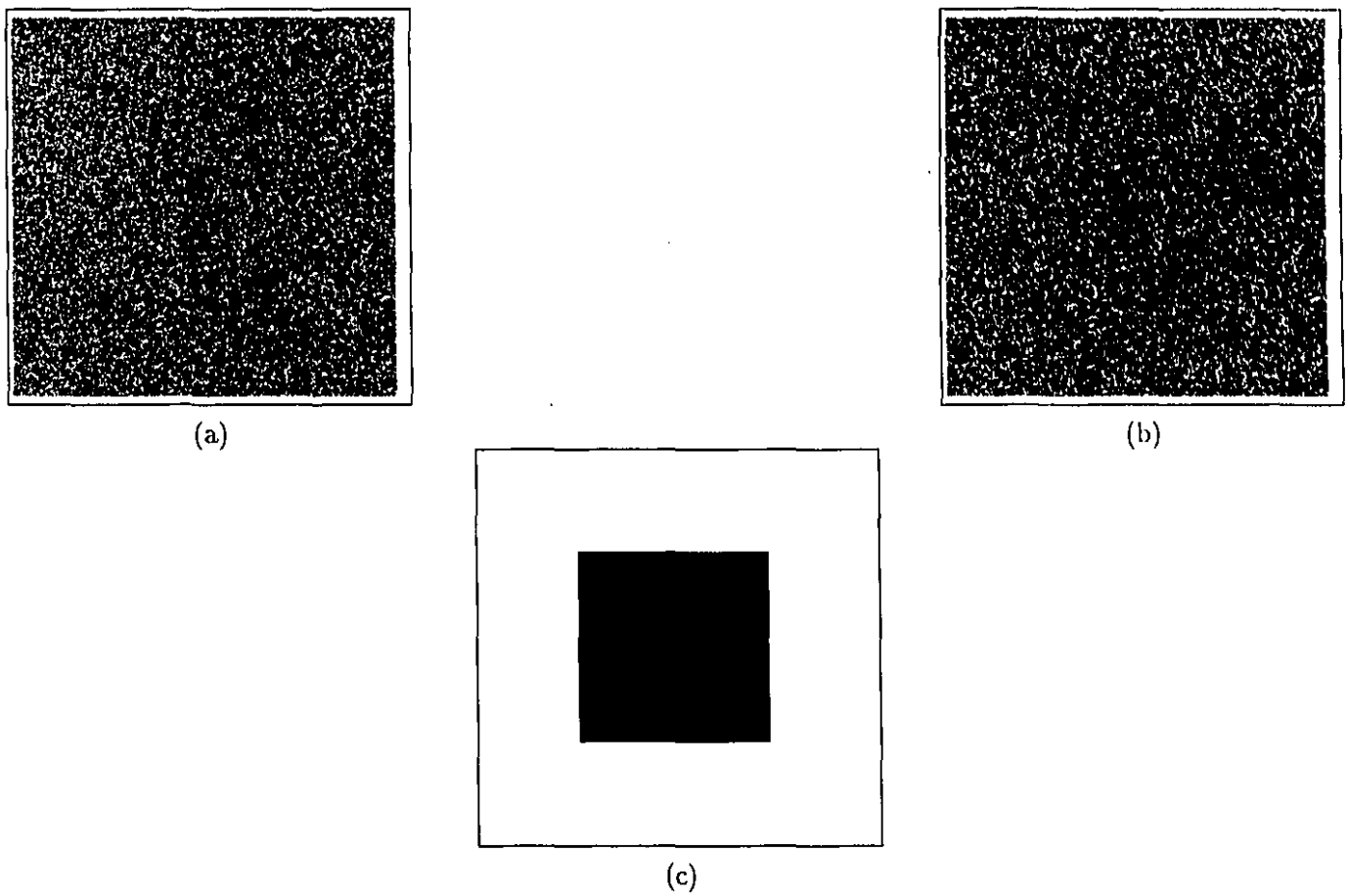


FIGURE 7.14. Random-Dot Stereogram. (a) Left image. (b) Right image. (c) Disparity map.

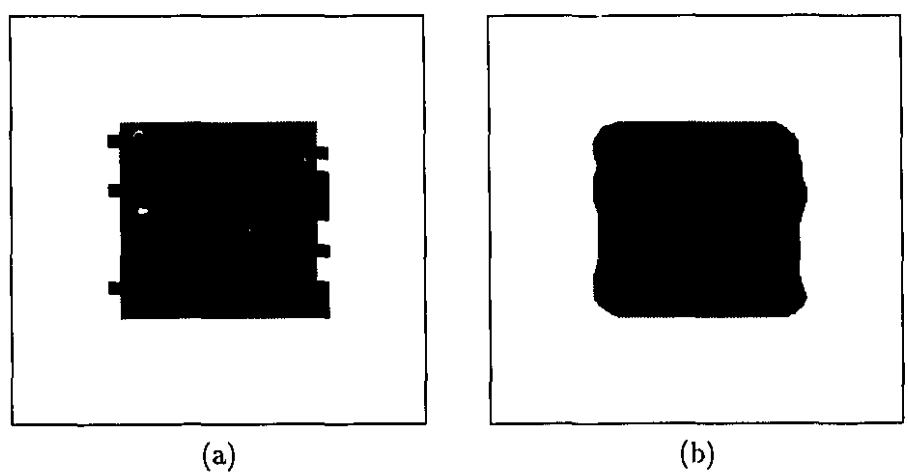


FIGURE 7.15. Performance of the Method on Random-Dot Stereogram. (a) Initial disparity map. (b) Refined disparity map.

and median filtering. The first two image pairs have been taken using aligned cameras and therefore do not contain a point or surface of fixation. As a result, surfaces that are closest to the camera have large disparities. Since disparity decreases with increasing distance, theoretically zero disparity corresponds to surfaces of infinite depth. However, in practice due to sampling considerations, surfaces that are located "far" from the imaging device have zero disparities.

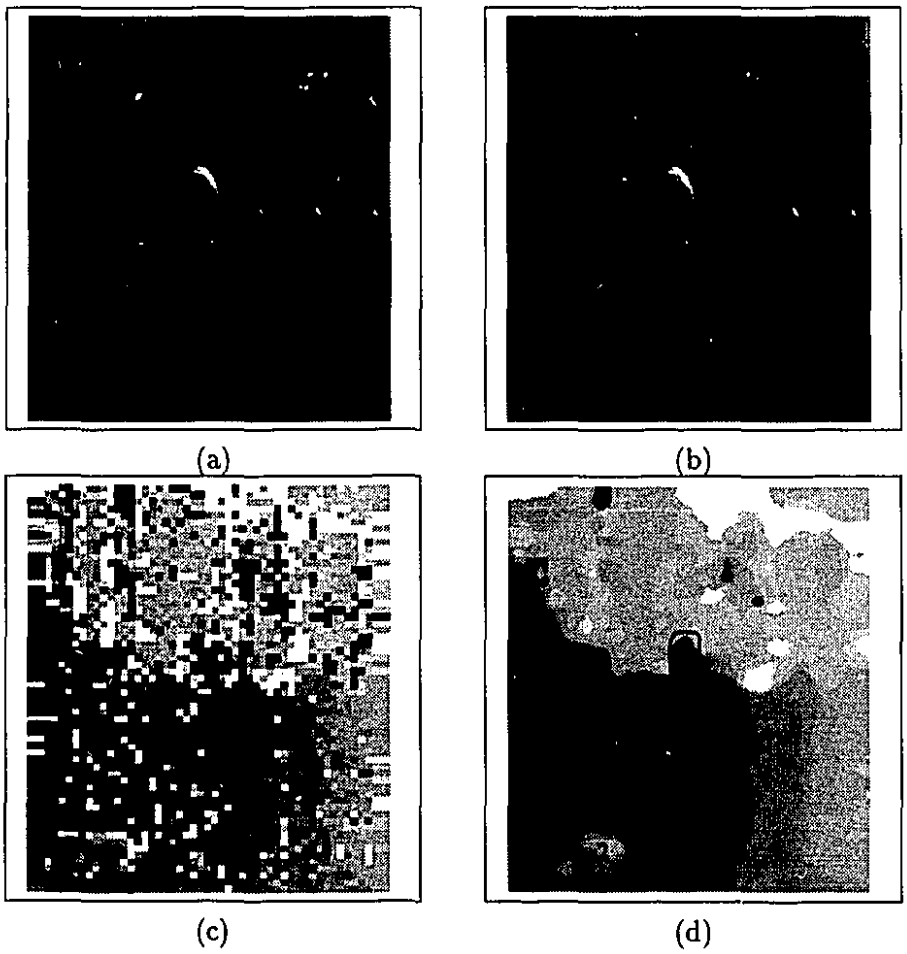


FIGURE 7.16. Performance of the Method on Real Stereo Images (Parking Meter). (a) and (b) Left and right images. (c) Initial disparity map. (d) Refined disparity map. 1 and 2 indicate the top of the parking meter and the opening in the shrubs respectively.

Consider the image pair in Figure 7.16. Despite the low level of texture in the original image pair, the algorithm has obtained the correct disparity map. The disparities of the shrubs and parking meters decrease with increasing distance from left to right. The disparity map even illustrates the presence of geometric features such as the top part of the meters

or the opening in the shrubs in the lower left part of the scene. A complete lack of texture on some parts of the building has caused areas of false disparity in this region in the initial disparity map. Refinement using median filtering has eliminated most of these regions.

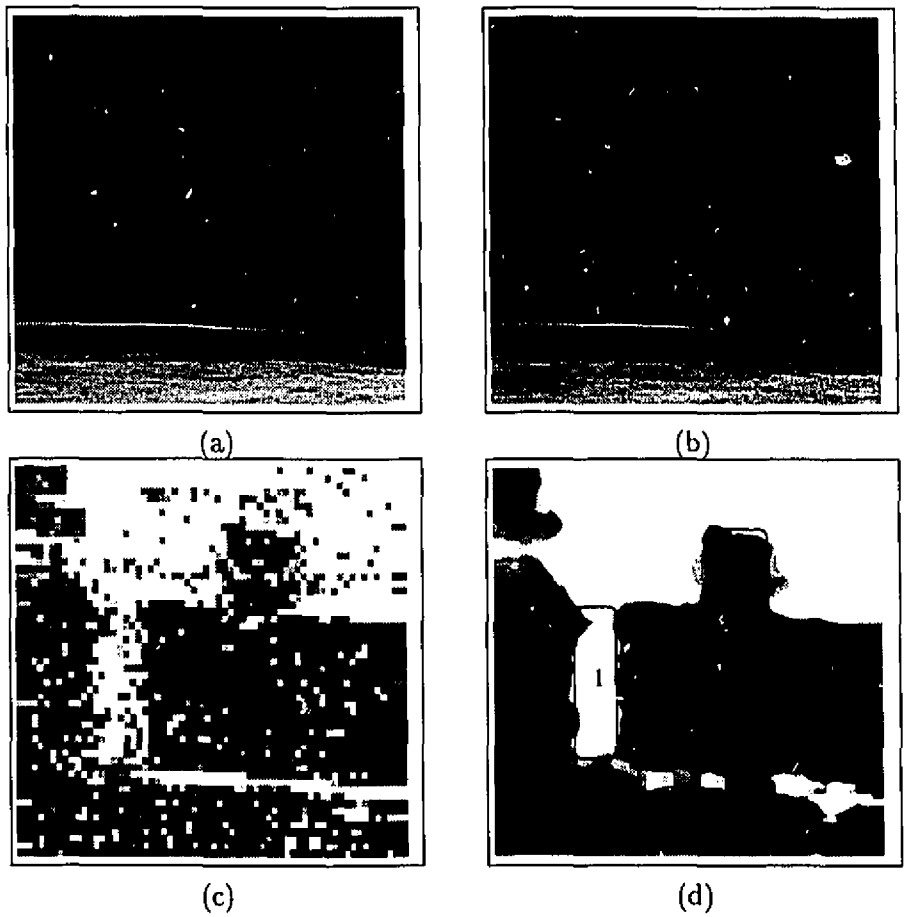


FIGURE 7.17. Performance of the Method on Real Stereo Images (Shrub). (a) and (b) Left and right images. (c) Initial disparity map. (d) Refined disparity map. 1 and 2 indicate the opening in the shrubs and the traffic sign respectively.

Now consider Figure 7.17. Once more the algorithm has correctly segmented The image into distinct regions based on the distance of surfaces in the scene. The false disparity in the white stripe below the shrubs corresponds to the low-textured surface of the curb. The disparity map is clearly appropriate for distinguishing objects - the traffic sign, for example - from the background. It also identifies landmarks such as the opening between the shrubs which can be used for path planning.

This example also illustrates the robustness of the cepstrum and the algorithm to differences between the two images of the stereo pair. There is a considerable difference of

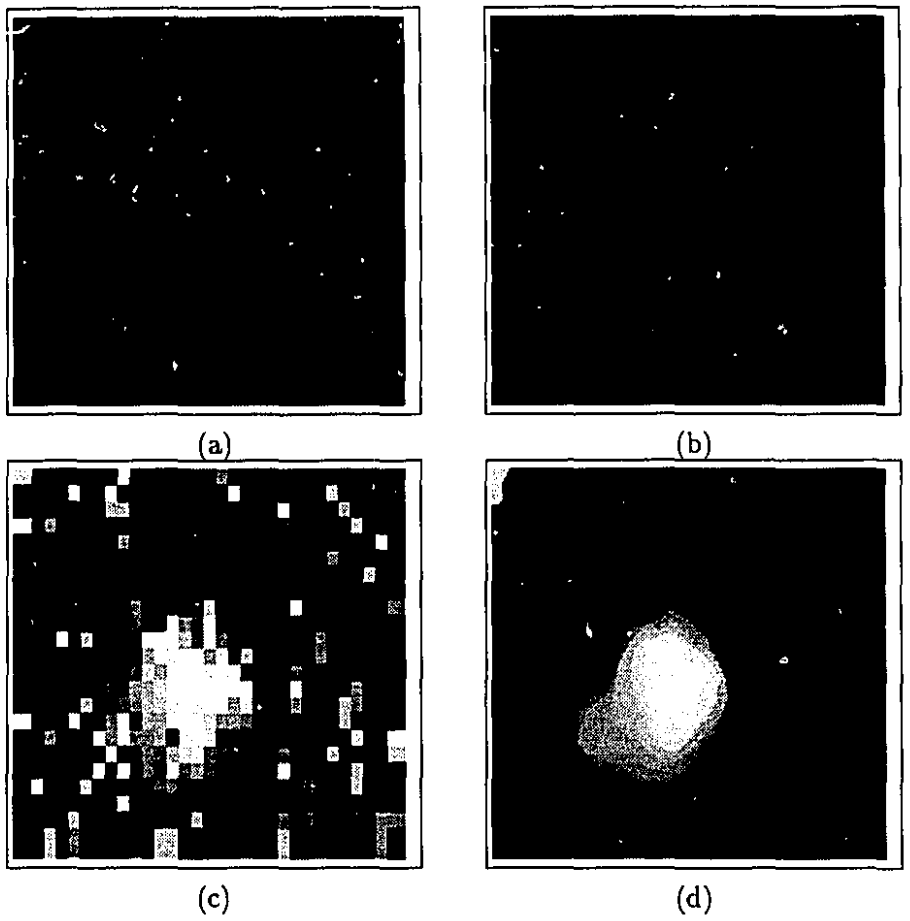


FIGURE 7.18. Performance of the Method on Real Stereo Images (Rock). (a) and (b) Left and right images. (c) Initial disparity map. (d) Refined disparity map.

5% between the brightnesses of the two images. Despite this, the segmentation provides the correct disparity map of the scene.

The final stereo pair used for experimentation with the algorithm is the image of a rock taken with the cameras verging on its surface. Thus there is an area of the surface where the disparity is very small. This image pair has been taken without camera calibration. Therefore, misalignment takes place not only because of vergence but also from the fact that the two cameras do not have the same vertical elevation. In addition, there is different lighting in each image and one of the two cameras is slightly out of focus. These factors typically exist in autonomous mobile robotic systems operating in an unstructured environment.

This example not only illustrates the robustness of the algorithm to the inter-ocular differences mentioned above but also provides an indication of the performance on surfaces with changing disparity. The final disparity map contains an indication of the boundary between surfaces at different distances. Further, it illustrates the variations in the surface of the rock itself.

CHAPTER 8

Biological Plausibility of a Correlation-Based Model

This chapter considers the plausibility of a correlation-based mechanism as a computational model for neural disparity estimation. The correlation is performed on the visual information in the ocular dominance columns of the primary visual cortex. It is important to distinguish between the two phrases "computational model of *binocular disparity estimation*" and the more commonly used "computational model of *stereopsis*". The former process results in a neuronal response to binocular disparity. This constitutes only a single stage in the latter, which is defined as encompassing the whole process of depth perception using vision through two eyes. This thesis not only makes no attempt to explain biological stereopsis but also points out that any such attempt, based on present knowledge of the visual cortex, is on shaky grounds. What the thesis does do is to associate the properties of neural disparity sensitivity in the early stages of the visual pathway with a computational model. Even then, we need to emphasize that we make no claims that this is the actual process which produces the response of these neurons. Rather, we simply state that the range of connections and properties of such neurons support such a model. Furthermore, as explained later, the term correlation does not refer to a specific operation. Instead, it is used to denote an "equivalent class" of operations which compare or *correlate* their two inputs.

There are a few occasions in this chapter where the analogies between the properties of neural disparity sensitivity - a single stage - and those of stereopsis - the overall process - are indicated. These similarities are also related to the computational model under study. After considering disparity estimation from a computational point of view, the chapter studies the plausibility of the overall disparity map.

1. Neural Mechanism of Tuned Neurons

1.1. The Connections Required for a Correlation-Based Mechanism of Neural Disparity Estimation. Assume that the dendrites of a complex Tuned Excitatory (TE) neuron have synaptic connections with the axons of the neurons in half of an ocular dominance column as well as the adjacent half of the adjacent column. The two ocular dominance columns of course correspond to different eyes. Due to the half-column overlap of the retinal area represented in neighbouring ocular dominance columns, the two half-columns contain information from *identical* locations in the two retinae. The disparity sensitive neuron receives its inputs from both of these half-columns. Such connections are schematically illustrated in Figure 8.1.

In this way, the two sets of monocular information are "combined" with each other. With such connections, the disparity sensitive neuron can act as a "filter" that provides a measure of the "correlation" between the information on the two retinae. In addition, the disparity selective neuron provides a medium to represent such correlation properties. The activation of a particular neuron by the correlation properties is dependent on the location of the peak of the correlation function and provides a mechanism for peak detection.

Such a model of a complex cell is different from the hierarchical model offered by Hubel and Wiesel in which a complex cell receives its input from a series of simple cells [**51**]. By contrast, the complex cells of layer 4B, such as that in Figure 8.1, receive their inputs from the neurons of layer 4Cα. In this layer, visual information is still in a circle-surround receptive field format and no simple cell structure is present. The disparity selective complex cell is activated when the "non-simple-cell" subfields in one eye are shifted versions of those in the other eye. Such a model of course requires a binocular or "dual" input and would not respond to a monocular input. With monocular stimulation, the subfields of one eye have zero input and spectrum. This is in agreement with the observations of Poggio *et. al*, stated in Section 2.2.1 of Chapter 3.

1.2. Estimation of Correlation in Neurons. Associating a specialized computation such as the cepstrum to cortical neurons is a far-reaching claim. Nevertheless, we observe that cortical neurons are believed to behave like spatial frequency filters and be involved in spatial frequency discrimination [**108**], [**120**], [**124**]. Such "frequency-based processing" would constitute the first step in estimating the correlation function. It is also important to note that spatial and frequency domain representations of signals are merely

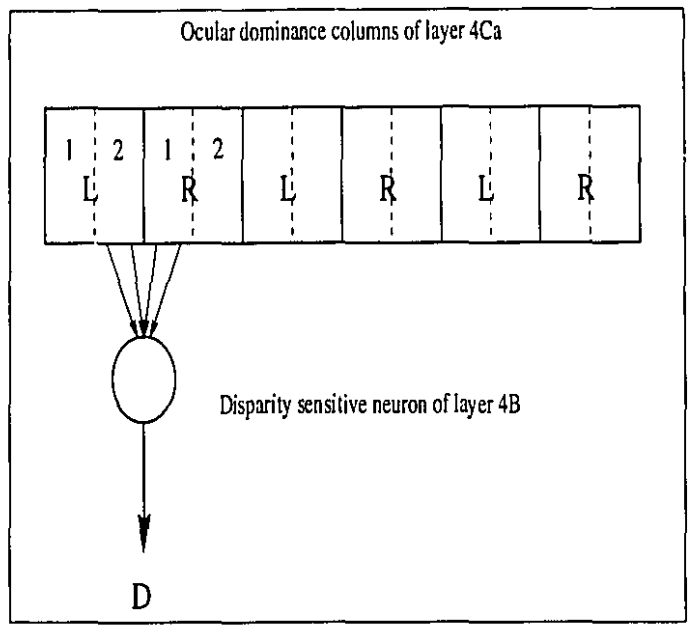


FIGURE 8.1. Neural Connections from the Ocular Dominance Columns to the Disparity Sensitive Neurons (schematic). L and R represent the ocular dominance columns corresponding to left and right eyes, respectively. The second half (marked by 2) of each column, corresponding to the left eye, has connections to the same neuron as the first half (marked by 1) of the adjacent column corresponding to the right eye, and vice versa. The two half-columns also represent the same retinal area (see Figure 3.2). The details of the mapping from the eyes to the ocular dominance columns are shown in Figure 3.1. The disparity sensitive neuron indicates the disparity between the two half-columns as its response. Disparity is measured by and represented in this neuron.

two representations of the same information. A processing unit such as a neuron is obviously not aware of the distinction between the different representations. As long as the response of the neuron to an "equivalence class" of inputs resembles the output of a process to the same class of inputs, the neuronal activity can be modelled by that process. Thus it seems that cortical neurons are involved in some form of frequency domain processing of visual data.

Even the above reasoning gives no indication of the particular "sub-class" of correlation operation performed between the information from the two eyes. From a computational point of view, it is only possible to state that a "correlation-like" operation may be the underlying factor for the response of Tuned Excitatory neurons. However, *correlation* is as specific a mechanism as the present neural knowledge can justify. Associating these cortical neurons with the cepstrum, a precise type of correlation algorithm, is too strong a claim considering the generality of the cortical computations.

1.3. Properties of Tuned Cells. This section explains some of the observed properties of the tuned cells using the ocular dominance column correlation-based model. The properties are fully explained in Chapter 3.

1.3.1. *Range of Detectable Disparities.* At any eccentricity, cortical complex cells have receptive field sizes approximately equal to the *overall* receptive field of (all the neurons in) one half of an ocular dominance column [**29**], [**51**]. With increasing eccentricity, the size of the receptive fields of these complex cells increases in the same manner as the retinal area represented in the ocular dominance columns. In the neuronal correlation disparity estimation mechanism, disparity sensitive neurons receive information from a region as large as one half of an ocular dominance column. With such an input, at any eccentricity the width of the ocular dominance columns determines the range of disparities detectable by the neuron. This can be compared with the maximum detectable disparity of the algorithm in Chapter 5.

The range of preferred disparities observed in the tuned neurons of layer 4B in fact falls within the visual angle subtended in one half of an ocular dominance column at all eccentricities. Furthermore, the range of preferred disparities of tuned neurons increases with increasing eccentricity in the same manner as the increase of the visual angle represented in ocular dominance columns.

1.3.2. *Width of Disparity Tuning Curve.* With a correlation model for disparity estimation, the "precision" of disparity estimation depends on the size of the subfields of the disparity selective neuron. Larger subfields result in coarser "sampling of the disparity space". As mentioned in Chapter 3, the receptive field size of individual neurons in layer 4Cα of the primary visual cortex increases with increasing distance from the fovea. At the same time, the number of receptive fields mapped into an ocular dominance column is believed to stay constant. Given such a structure, the correlation mechanism predicts a decrease in the precision of stereo tuning. This decrease is represented by increasing width of the response profile of the disparity selective neuron at larger eccentricities. Furthermore, if disparity estimation in such neurons is indeed the initial stage of stereopsis, the increased "coarseness" of disparity estimation may in turn result in decreasing stereo acuity with increasing eccentricity. This is indeed true for human stereopsis.

1.3.3. *Global Stereopsis.* Before experiments with random-dot stereograms, the mechanisms of disparity sensitive neurons were explained as "matching" the stimuli on two receptive fields located at different retinal locations in the two eyes. However, no mechanism

for performing this matching operation was offered. The need for explanation became more apparent when it was discovered that some disparity neurons were actually responsive to random dot patterns. The disparity sensitive neurons seem to signal the correct disparity without ambiguity despite the numerous identical dots and their random placement in such patterns.

The neurons of the primary visual cortex receive inputs from a limited neighbourhood and, especially at the initial processing level, global interactions do not exist. Therefore, a global solution to the problem of false matches of dots in a random pattern, at a level as early as layer 4B, does not seem likely. A correlation-based model, by eliminating the need for matching individual subfields, eliminates the problem of false matches in a random dot pattern.

1.3.4. *Distinction between Tuned Zero and Tuned Far or Tuned Near Neurons.* Tuned Far (TF), Tuned Near (TN), and Tuned Zero (T0) neurons were introduced in Chapter 3 and illustrated in Figure 3.3. Using a correlation model, the Tuned Far and Tuned Near cells can be described as those neurons detecting the correlation peaks located at larger disparities (within the detectable range). Such disparities are near the "end" of the range spanned by the neuron where the correlation function extends towards zero from its peak location, but not in the opposite direction. Larger disparities lie outside the range of shifts represented in the given ocular dominance column. The correlation function, which in essence looks like that of a Tuned Zero curve, then "loses" its trailing edge towards larger shifts. The latter are not represented in the range and only the trailing edge towards zero remains.

1.3.5. *Opposite Contrast Stimuli.* A correlation-based algorithm involves the power spectral properties of the visual data which causes the sign of the stimulus contrast to be neglected. However, disparity sensitive neurons, and particularly those sensitive to binocular correlation, do not respond to stimuli having opposite contrast on the two eyes. This can be related to the existence of the two ON and OFF pathways in the visual system, each responsive to data of a given contrast *sign* [**64**], [**103**], [**111**], [**112**], [**127**]. Separate processing of the information in the two pathways guarantees that activation is in response to stereo stimuli with the same contrast sign.

1.4. Other Disparity Sensitive Neurons. Tuned Inhibitory neurons can be modelled in an identical manner to the TE neurons. The neuronal response is merely suppressed, rather than facilitated in the presence of the preferred disparity. On the other hand, Far and

Near Neurons respond to a much larger range of disparities than the Tuned neurons at any disparity. The application of the correlation model to Reciprocal Neurons requires a much larger "spreading out" of the dendritic connections of the Far (FA) or Near (NE) neuron, which is not a characteristic of many visual cortical cells. However, Reciprocal Neurons are observed more in layers beyond layer 4B of V1 and their response is likely to be due to a different mechanism. Simple disparity selective cells which are rare in layer 4B are also likely to share a similar situation. FA and NE neurons may be involved in determining the relative position of objects even when the images on the two eyes are not fused and diplopia occurs. Finally, the tuned neurons of other cortical layers and areas may receive their input from the tuned neurons of layer 4B.

2. Properties of Stereo Perception

2.1. Inter-Ocular Correlation. Chapter 3 stated that, for a given decrease in inter-ocular correlation, the stimulus contrast must be increased by squaring to maintain stereopsis. This property is supported by the correlation model. In this model, neural activity is a representation of the magnitude of the correlation function at the preferred disparity. Maintaining the same level of activity requires an increase in power content if the correlation is reduced.

2.2. Figure-Ground Separation Versus Absolute Depth Perception. The overall disparity map, obtained after cepstral disparity estimation and the subsequent refinement stage, contains basic information about the properties of each surface in the scene. However, its most salient features are the boundaries between surfaces. This is similar to the properties of human stereopsis explained in Section 4 of Chapter 3.

2.3. Imprecision of Stereopsis. In our disparity estimation algorithm, the presence of estimation noise limits the accuracy of the sub-pixel disparity by no more than one half of a pixel. Also, assigning a single disparity to the whole ocular dominance column may cause a shift in the boundary between surfaces. As mentioned in Section 4 of Chapter 3, imprecision also exists in human stereopsis. Besides lack of accuracy in perception of absolute depth, the perceived *depth difference* between surfaces is also imprecise [**76**]. This property, along with the importance of surface boundaries, motivates the notion that the method presented in this thesis, as well as human stereopsis, are tools for figure-ground separation.

CHAPTER 9

Implementation

Computational efficiency is the underlying principle for implementing the method we have developed for obtaining the disparity map of a stereo image pair. Local estimation of disparities permits the parallel implementation of the algorithm. This, in turn, increases the processing speed. With local estimation, a fraction of the image as small as a single ocular dominance column may be processed independently from the remainder of the image. Therefore, the complete implementation scheme may comprise as many parallel streams as the number of ocular dominance columns. Indeed this would resemble the organization of the stereo disparity "computation" in the primary visual cortex.

Parallel implementation is also favoured by the fixed running time of all levels of the algorithm. Fixed running time eliminates the need for taking into account the fact that different processors may receive different fractions of the task depending on the properties of the processed image. It permits the division of the overall processing task or data among multiple processors.

In addition to parallel implementation, two other strategies have been utilized which contribute to increasing the processing speed of of the algorithm. First, the method developed in the previous chapters is slightly modified to provide a considerable increase in the processing speed without compromising the performance of the algorithm or its usefulness for recognition purposes. Second, some of the characteristics of the algorithm which can reduce the processing time are exploited. These strategies are explained in the next two sections followed by the details of the parallel implementation.

1. Modifications to the Method

The discussions and experiments of Chapters 4 to 7 provide insight into the role of the factors which influence the performance of the algorithm. We exploit this knowledge and provide simplifications to the algorithm which reduce its computational complexity without

significant effects on its performance or use. The following sections provide a description of each of these modifications and compare any new output to those of Chapter 7.

1.1. Compensating for DC Signal Correlation While Re-Scaling the Cepstrum. Autocorrelation can be expressed as the sum of the products of the samples of a signal and its shifted version. For an uncorrelated signal, as shown in Figures 4.7 and 4.8, only the corresponding samples of the signal and its shifted version which occupy identical locations contribute to the autocorrelation sum. The sum of the products of the remaining samples is zero. For a correlated signal, on the other hand, there are also contributions from the non-corresponding values. By definition of signal correlation, although these values do not correspond, the sum of their products is non-zero. This additional contribution to autocorrelation is of course reflected as a smaller decrease in the magnitude of the autocorrelation peak as disparity increases when compared to those given by Equations 4.13 and 4.14. The slower decrease in the magnitude of the autocorrelation peak in turn influences the magnitude of the cepstral impulse by making the decrease of the latter more moderate.

For the specific case of DC correlation, the slower rate of decrease in the magnitude of the autocorrelation or cepstral peak can also be explained in the following manner. The presence of the DC component causes a vertical shift in the autocorrelation function or the cepstrum of the signal. This addition of a constant to the autocorrelation or the cepstrum increases the smaller values by a greater factor *relative* to the larger ones. Equivalently, for positive a, b, and c, if

$$\frac{a}{b} > 1 \tag{9.1}$$

then

$$1 < \frac{a+c}{b+c} < \frac{a}{b} \tag{9.2}$$

Of course, the rate of change of the peak magnitude with (increasing) disparity is indicated by the slope of the peak magnitude-versus-disparity curve. Therefore, moderating the decrease in the magnitude corresponds to flattening the peak magnitude-versus-disparity curve. In other words, signal correlation results in a slower decrease of the cepstral peak with increasing disparity and a flatter scaling curve. The exact degree to which the magnitude-versus-disparity curve is flattened depends on the value of the DC component of the signal.

However, regardless of the specific value and in common with all signals having non-zero mean, the decrease is more moderate than for uncorrelated signals.

Removing the signal mean eliminates the factor which moderates the decrease and makes the use of the re-scaling factors of Equation 5.2 appropriate. Data independence of the initial estimation stage is a primary concern in increasing the computational speed of the algorithm. Therefore, instead of computing and removing the mean of every image window, it is desirable to use a strategy which is independent of the processed signal. For example, one can choose a re-scaling factor that inherently represents the reduced rate of decrease in the cepstral peak magnitude for disparities greater than zero. This can be achieved by using the (inverse of) the ratio described by Equation 9.3 instead of that in Equation 5.2.

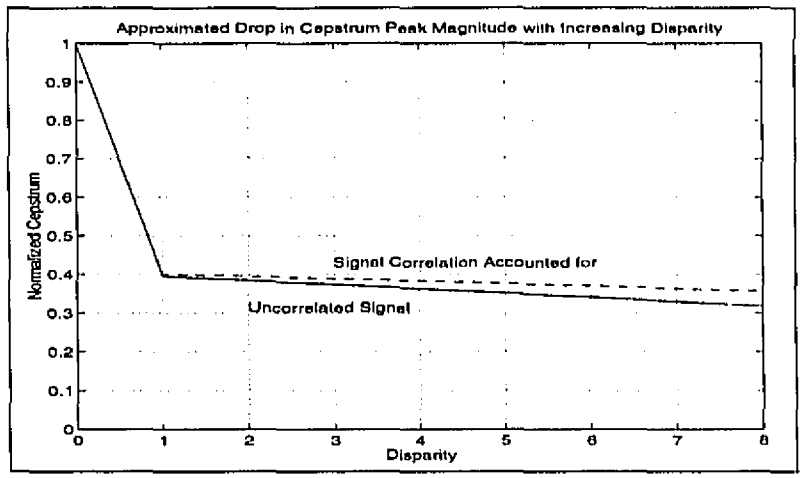
$$\frac{Peak_{ceps}(d)}{Peak_{ceps}(0)} = \begin{cases} 1 & d = 0 \\ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{\left(\frac{W-d}{2W-d}\right)^m}{1^m} = log\left(\frac{W-d}{2W-d} + 1\right) & d \neq 0 \end{cases} \tag{9.3}$$

Replacing the factor $2W$ by $2W - d$ in the (autocorrelation) ratio corresponds to influencing both the numerator and denominator of this relationship with the increasing disparity factor, rather than just the former. Note that because the term d is subtracted from W in the numerator and from $2W$ in the denominator, there is still a decrease in the magnitude with increasing disparity. However, this decrease is at a slower rate because of the new term in the denominator. This, in turn, corresponds to making the peak magnitude-versus-disparity curve flatter, as illustrated in Figure 9.1 (a).

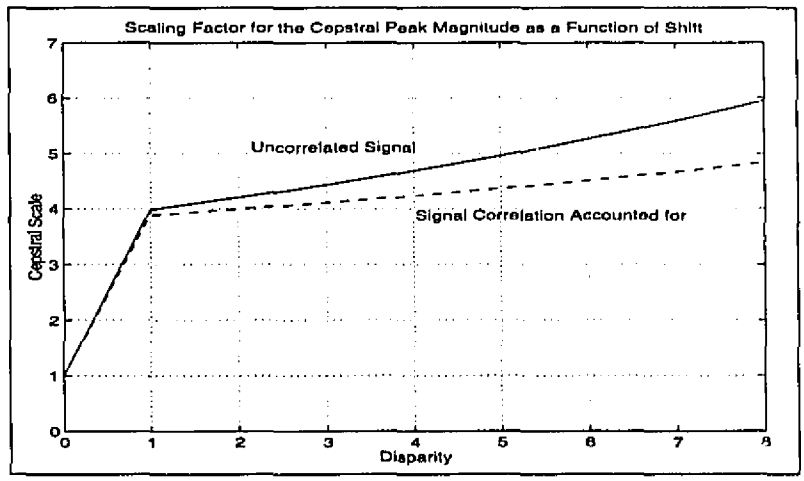
The presence of the signal mean shifts the cepstrum vertically. Therefore, the cepstrum of the window is in general larger than that of the original method. To account for this increase in the magnitude, the cepstral threshold of Chapter 6 is also increased. Specifically, the threshold is doubled from its original level of $1\% - 2\%$ of the value $W \times H \times \log(W \times H)$ to $2\% - 4\%$ of this value.

Figure 9.2 illustrates the output of the algorithm when using the new scaling factors and corresponding thresholds for some of the test images used in Chapter 7. The results are nearly identical to those of Figures 7.10 and 7.17 where the original method of removing the mean and scaling the cepstrum was used.

1.2. Disparity Estimation with Pixel Accuracy. Disparity estimation with pixel accuracy is used for the implementation of the fast version of the algorithm. Abandoning the use of sub-pixel accuracy for disparity estimation in favour of pixel accuracy increases the processing speed in two ways. First, it reduces the time required for disparity estimation



(a)



(b)

FIGURE 9.1. Accounting for Signal Correlation in the Cepstral Re-Scaling Factor. (a) The decrease in the cepstral peak for uncorrelated and correlated signals. The curve for the uncorrelated signal is from Figure 5.1. The curve labelled "signal correlation accounted for" considers signal correlation and is therefore flattened. (b) The corresponding re-scaling factors. The re-scaling factors are the inverse of the decrease in the cepstral peak given in part (a). The length of the original signal is 32 samples and signals are represented by concatenation.

by eliminating the search for the largest neighbour of the cepstral peak and and also the ensuing interpolation between the two. Second, pixel accuracy disparity estimation reduces the computations required for refining the disparity map. Maintaining sub-pixel accuracy in the refinement stage requires computing the median using a sorting scheme. Discrete

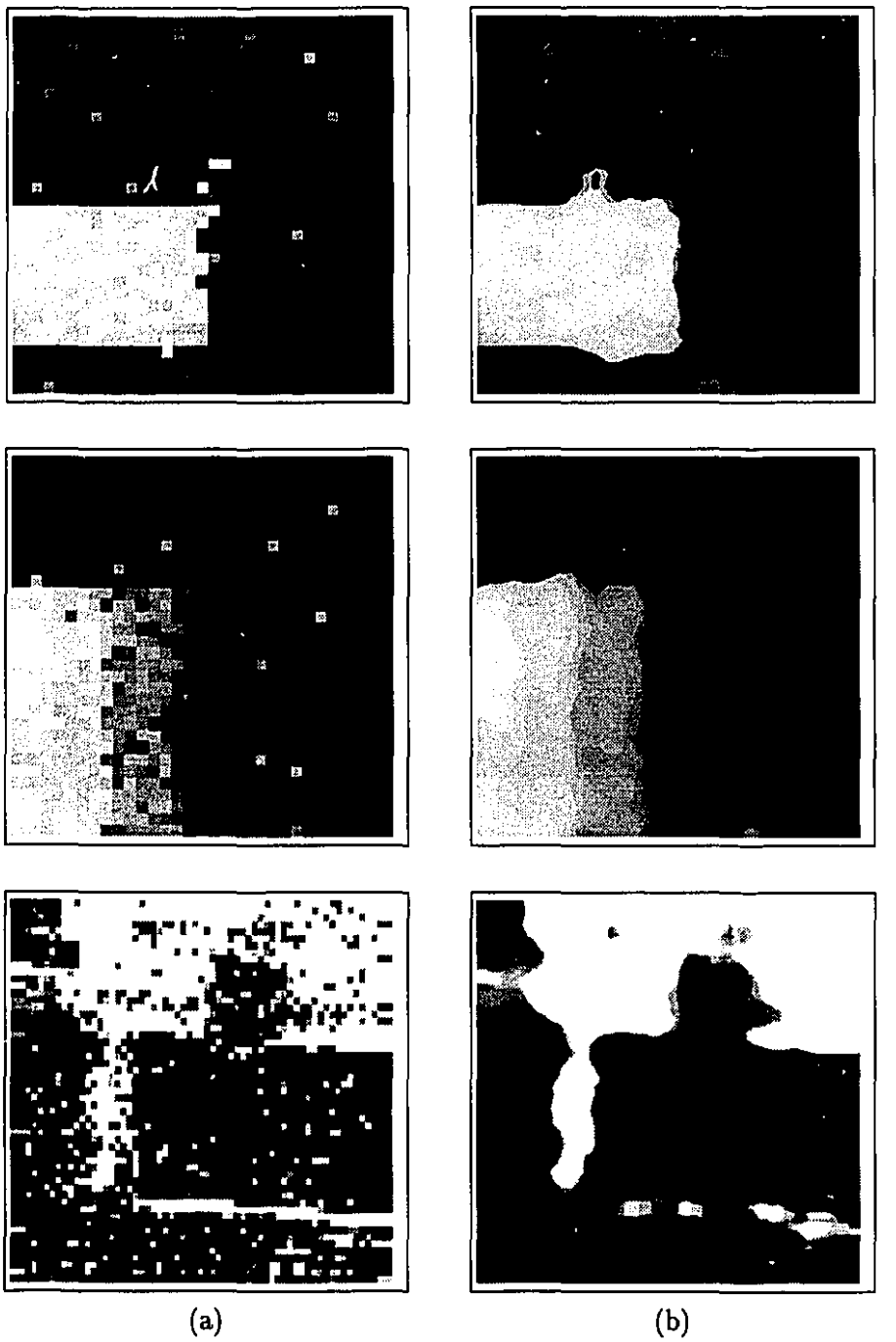


FIGURE 9.2. Experimental Results with Compensation for Signal Correlation While Re-Scaling the Cepstrum. (a) Initial disparity maps. (b) Refined disparity maps.

disparities permit the use of a fast median filtering method of [47] which offers a considerable increase in the speed of the filtering process.

Pixel accuracy disparity estimation reduces the ability of the method to represent surfaces whose disparities change. However, it does not affect its ability to detect those places where a *sudden* change in disparity occurs. Therefore, the shift from sub-pixel to pixel accuracy does not undermine the method's ability to discriminate between surfaces of different depth. In a system that uses stereopsis for figure-ground separation, implementation of disparity estimation to the nearest pixel provides considerable computational savings. The experiments of Sections 1.1.4 and 1.1.6 of Chapter 7 illustrate the outcome of the algorithm with both approaches to the disparity estimation process.

1.3. Data Domain Filtering. As mentioned in Section 3.2 of Chapter 6, and shown by the experiments of Section 1.2.2 of Chapter 7, median filtering in the image domain results in smoother boundaries between regions of different disparity. This added advantage is provided at the cost of a considerable amount of extra computation. The number of points filtered using image domain filtering is two orders of magnitude greater than the number of filtered disparities in the data domain for typical ocular dominance column dimensions. This is because the former equals the total number of pixels in the image compared to the number of ocular dominance columns, which defines the latter.

Furthermore, in the data domain, each ocular dominance column is represented by a single disparity whereas in the image domain as many values as the number of pixels in the column are used in its representation. Therefore, if the size of the median filter in units of visual angle is the same in both schemes, the image-domain filter contains a larger number of samples. This is the second factor which causes an increase in computation for filtering in the image domain.

For the above reasons, filtering in the data domain offers an added advantage of considerable computational savings and is adapted for parallel implementation. Once again, the slight roughness of disparity region boundaries associated with data domain filtering does not undermine the usefulness of the result for figure-ground separation. Figure 7.11 (b) illustrates the effect of data-domain refinement.

2. Implementation Considerations

The computation of the two-dimensional cepstrum can be divided into the following steps:

- computing the two-dimensional Fourier-transform of the signal.

- obtaining the magnitude of the Fourier transform or its square.
- performing a logarithmic transformation on the amplitude or power spectrum of the signal.
- taking the forward or inverse Fourier transform of the log spectrum.
- obtaining the magnitude of the second Fourier transform.

Using the definition of the Fourier transformation, a two-dimensional discrete Fourier transform can be divided into two interchangeable sets of one-dimensional Fourier transforms. The one-dimensional transforms are performed along the rows and columns of the transformed matrix. Therefore, the implementation of the algorithm involves a set of row and column Fourier transforms. This is then followed by taking the logarithm of the complex magnitude, another set of row and column Fourier transforms, and finally a complex magnitude. The following sections explain specific considerations which reduce the total computational effort required for the implementation.

2.1. Reduction in the Number of the Fourier Transforms. Overlapping ocular dominance columns permit a reduction in the number of one-dimensional Fourier transforms performed for obtaining the first two-dimensional Fourier transform of the concatenated image window. Adjacent image patches share as many columns as the the extent of overlap between them. If Fourier transformation along columns precedes transformation along rows, the shared columns need to be transformed only once. This provides considerable savings in computing the Fourier transformation. For example, with an overlap of 50% the number of the first set of transformations along the columns is nearly halved.

It is also possible to reduce the number of transformations performed to obtain the second two-dimensional Fourier transform. Since peak detection is performed in a fraction of the cepstrum of the concatenated window, the values of the remaining parts are of no use. Assuming that in the second two-dimensional Fourier transform the transformation of columns follows that of rows, only those columns that are involved in peak detection need be transformed [**85**].

Including the strategy for reducing the number of the second set of column transforms into the implementation is a trivial task. However, that of the first set requires special implementation considerations. For the first set, after transformation along columns each image window undergoes row transformation. To re-use the values of the shared columns, the implementation should ensure that these columns are stored and available for use with the next window.

2.2. Reduction in the Computational Effort Required for the Complex Magnitude Calculation. As explained in Chapter 4, the cepstrum can be defined using both the amplitude and power spectra of the signal. The two definitions correspond to performing the logarithmic transformation on the complex magnitude of the Fourier transform or the square of this value. The logarithms of the two values differ only by a factor of two.

The magnitude of any complex number z is defined as $\sqrt{(Re(Z))^2 + (Im(Z))^2}$. In practice, calculating the amplitude spectrum is computationally more expensive than the power spectrum because of the square-root operation involved in obtaining the complex magnitude. To avoid the additional computation, the implementation of the method uses the power spectrum of the signal. To account for the doubling of cepstral values, the cepstral threshold is also doubled.

The Fourier transform of a real signal is an even function of frequency. As a result, the power spectrum of such a signal is an even and real function whose Fourier transform is in turn real. Therefore, the power cepstrum of a real signal is also real, making the computation of the second complex magnitude unnecessary. The implementation of the algorithm employs this property. It uses the value of the cepstrum directly in peak detection and avoids the unnecessary computation of obtaining the complex magnitude.

2.3. Substituting the Hartley Transform for the Fourier Transform. The Hartley transform of a real sequence $f(m)$, $m = 0, 1, \cdots, M-1$ is defined as:

$$H(u) = \sum_{m=0}^{M-1} f(m) \left(\cos\left(\frac{2\pi um}{M}\right) + \sin\left(\frac{2\pi um}{M}\right)\right) \tag{9.4}$$

The Fourier transform of the same sequence is given by:

$$F(u) = \sum_{m=0}^{M-1} f(m) \left(\cos\left(\frac{2\pi um}{M}\right) + j \sin\left(\frac{2\pi um}{M}\right)\right) \tag{9.5}$$

In computing the cepstrum of a real signal, the Fourier transform can be replaced by the Hartley transform [**114**]. Using the Hartley transform instead of the Fourier transform for obtaining the cepstrum offers a reduction in computational effort as well as memory requirements. This increased efficiency is mainly due to the fact that the Hartley transform does not involve an imaginary component. The authors in [**114**] claim that the use of the Hartley transform for computing the cepstrum reduces the computational time and memory requirements by approximately 40% and 50%, respectively. It should be noted that the exact

reduction in the computational effort is strongly dependent on the specific algorithm used for obtaining the Fourier and Hartley transforms of a sequence. The computational savings also depend on the time involved in computing the sine and cosine values used in the two transformations.

Despite this, any implementation of the Hartley transform does offer computational savings for computing the cepstrum compared to a similar implementation of the Fourier transform. In addition to the direct computational savings, the reduced memory requirement results in additional reduction in the processing time within the specific implementation circumstances of this work. A smaller memory requirement allows the use of on-chip memory for the processing buffers which reduces the memory access and the total processing time. This aspect is fully discussed in Section 3.3.

The reduced memory requirements of the Hartley transform as compared to the Fourier transform is an illustration of the manner in which the two schemes store information. In the Fourier transform, the even and odd parts of the transformed sequence - represented by cosine and sine components, respectively - are stored *separately* as the real and imaginary part of this sequence. In the Hartley transform, this same information is stored as the even and odd components of the *same function.* Therefore, the latter method conserves half of the memory required for information storage. The cost of such conservation is that the even and odd parts of the elements of the sequence are not directly accessible. In general, to retrieve the even or odd part of each element, the element itself, as well as others, need to be used. Consequently, despite the fact that no imaginary parts are computed, the saving that results from the Hartley transform is *less* than 50% of the total computation required for obtaining the cepstrum from the Fourier transform. Retrieving the even and parts of each element from the whole sequence can easily be achieved if we note that any function of one variable, $f(u)$, can be described in terms of an even and an odd component. These are obtained from the following two relationships [**62**]:

$$\begin{aligned} f_{even}(u) &= \frac{f(u)+f(-u)}{2} \\ f_{odd}(u) &= \frac{f(u)-f(-u)}{2} \end{aligned} \tag{9.6}$$

For a finite length discrete-time signal $f(m)$, $m = 0, 1, \cdots, M-1$, the Hartley transform is periodic with a period of M. In other words, for such a signal, $H(-u) = H(M-u)$. Using this relationship and Equation 9.6 to obtain the even and odd components of a sequence, one can obtain the power spectrum as

$$
\begin{aligned}
| F(u) |^2 &= (Re[F(U)])^2 + (Im[F(u)])^2 && u = 0, 1, \cdots, M-1 \\
&= F^2_{even}(u) + F^2_{odd}(u) && u = 0, 1, \cdots, M-1 \\
&= H^2_{even}(u) + H^2_{odd}(u) && u = 0, 1, \cdots, M-1
\end{aligned} \tag{9.7}
$$

or equivalently

$$
\begin{aligned}
| F(u) |^2 &= \begin{cases} H^2(0) & u = 0 \\ \frac{1}{4}\left((H(u) + H(M-u))^2 + (H(u) - H(M-u))^2\right) & u = 1, \cdots, M-1 \end{cases} \\
&= \begin{cases} H^2(0) & u = 0 \\ \frac{1}{2}\left(H^2(u) + H^2(M-u)\right) & u = 1, \cdots, M-1 \end{cases}
\end{aligned} \tag{9.8}
$$

Appendix D illustrates that, for a two-dimensional signal and a two-dimensional Hartley transform, the even and odd parts of the signal are each stored with two-dimensional symmetry. There is also a one-dimensional symmetry between them. With such a relationship, the power spectrum of the signal is obtained from its two-dimensional Hartley transform, $H(u, v)$ as shown in Equation 9.9 as follows:

(9.9)

$$
\begin{cases}
| F(0,0) |^2 = (H(0,0))^2 & \\
| F(u,0) |^2 = \frac{1}{2}\left(H^2(u,0) + H^2(M-u,0)\right) & u = 1, \cdots, M-1 \\
| F(0,v) |^2 = \frac{1}{2}\left(H^2(0,v) + H^2(0,N-v)\right) & v = 1, \cdots, N-1 \\
| F(u,v) |^2 = \frac{1}{4}\Big((H(M-u,v) + H(u,N-v))^2 & \\
\qquad + (H(u,v) - H(M-u,N-v))^2\Big) & u = 1, \cdots, M-1; v = 1, \cdots, N-1
\end{cases}
$$

Once again, the extra computation required for obtaining the power spectrum of the signal from its Hartley transform reduces the savings which actually result from substituting the Fourier transform by the Hartley transform. Nonetheless, the substitution still provides a decrease in the total computational effort. Also the reduction in the required memory helps to decrease the processing time. The Hartley transform is performed using the Fast Hartley Transform (FHT) algorithm of [**20**].

3. Parallel Implementation of the Algorithm

The algorithm was implemented on a network of TMS320C40 processors as a part of a more elaborate vision system. A thorough description of the features of the processors is

provided in [19]. The processor features most attractive to the parallel implementation of the method developed in this thesis are the inter-processor communication capabilities and the direct-memory-access (DMA) co-processor. The extensive communication capabilities of TMS320C40 processors permit data transfers between distinct processors without any need for intermediate hardware. The DMA co-processor complements the communication capabilities by undertaking the transfer of information between memory and communication ports and freeing the central processing unit (CPU) from this task. The DMA co-processor can also perform memory-to-memory transfer and prepare data for processing without any computational load on the CPU. This allows implementation schemes which process data at specific memory locations without the computational cost associated with transferring information to such locations.

3.1. Data Distribution Versus Task Distribution Implementation. Multiprocessor implementation can lie on a spectrum whose extreme points are marked by the two schemes of data distribution and task distribution. In a task distribution or pipeline implementation all processors operate on *all of the processed data* but perform *disjoint operations* on the information. In a data distribution implementation, on the other hand, different processors perform *the same operation* on *different segments of the overall data.* It is possible to combine the two schemes and use multiple processors in a variety of manners depending on the format of the overall task and data.

Given a fixed number of processors, the throughput and latency of the system are the other factors which determine the implementation strategy. System throughput is defined as the rate at which information is processed by the system and is measured in units of data per unit time. Latency is the time elapsed between the arrival of one unit of data at the system and the exit of the result. It is measured in units of time. Although higher throughput and lower latency are always favoured, specific requirements, and especially the balance between the two, is dependent on the underlying task.

The throughput of a system with sequential components equals the throughput of its slowest one. The latency of such a system equals the sum of latencies of all its components. For a parallel combination of processors, preceded by a distribution stage and followed by a recombination stage, the throughput and latency are both determined by the slowest component. Although throughput and latency are often related, the exact relationship is dependent on the specific configuration of system. Changing a component of the system may influence both, one, or none of these parameters.

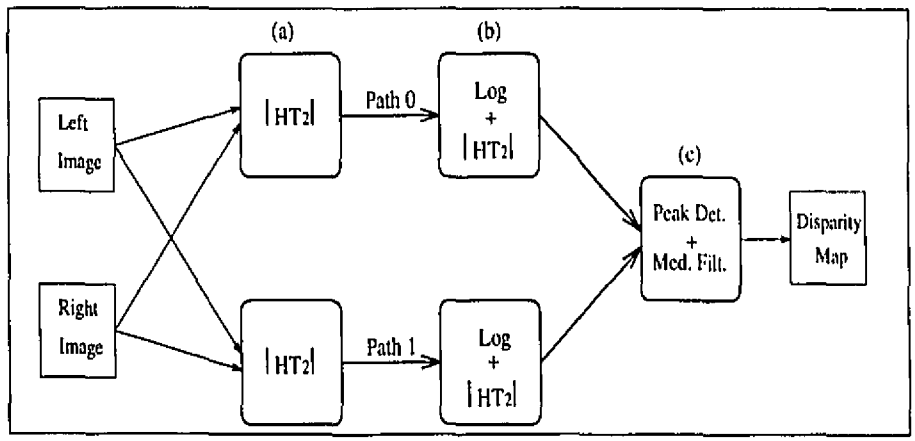


FIGURE 9.3. Parallel Implementation for Obtaining the Disparity Map of a Stereo Image Pair. (a) The first processor of both paths computes a two-dimensional Hartley transform and obtains the power spectrum from it. (b) The second processor obtains the log spectrum and performs the second two-dimensional Hartley transform to compute the cepstrum. (c) The recombination processor performs peak detection on the cepstrum, combines the disparities of both paths, and refines the overall disparity map. More processors would reduce the processing time in a linear manner.

3.2. Implementation Scheme. The disparity map computation in this thesis is implemented on five TMS320C40 processors. It involves a combination of data and task distribution schemes, as shown in Figure 9.3. In addition to specific implementation considerations explained next, the implementation scheme of Figure 9.3 makes the extension to more parallel pipelines of processors a trivial task. Since disparity estimation is performed independently for each ocular dominance column, the image pair can be divided into more parallel pathways using more processors. Furthermore, using more processors, the processing requirements of each stage can also be divided more finely. The processing speed will then increase in as a linear function of the number of processors.

The overall task of computing the cepstrum can easily be divided into two approximately equal "sub-tasks". Because of the reduction in the number of column transforms for optimization purposes, the division of the task into more than two *distinct and equal* tasks for a complete pipeline implementation is not possible. Furthermore, fewer sequential processors, along with the presence of parallel pipelines, reduce the latency of system. This is the motivation for implementing pipelines that each contain two processors.

The recombination processor uses some of its available DMA channels for internal memory-to-memory communication. This leaves a limited number of channels for receiving input data from external processors. Therefore, parallel pipelines of two processors

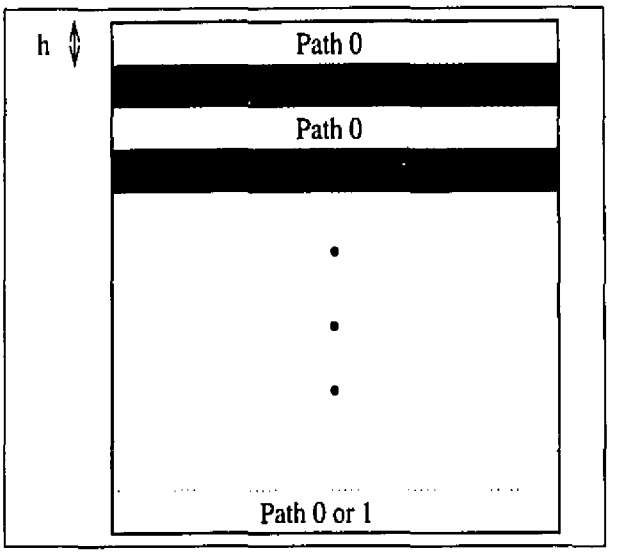


FIGURE 9.4. Division of the Stereo Image Pair for Parallel Processing. h is the height of an ocular dominance column. Both images of the stereo pair are divided according to this scheme. The two parallel pipelines process alternating stripes of the image pair. The first path may process an extra stripe depending on whether the total number is odd or even.

each are preferred to a completely parallel configuration for computing the cepstrum. The latter demands additional communication channels from the recombination processor. In addition, data transfer links with many parallel processors would use some of the internal processing resources of the recombination node. This is despite the fact that the DMA co-processor bears the responsibility of data transfers required for inter-processor communication. Therefore, the current architecture is more efficient than performing the task of computing the cepstrum by a completely parallel configuration.

The reduction in the number of column transforms, motivated by the presence of overlapping patches, requires all adjacent columns to be processed by a single pipeline. Therefore, each image of the image pair is divided into "stripes" whose height equals that of the ocular dominance columns. Alternating stripes are passed through and processed by the two paths. This division scheme is shown in Figure 9.4.

The first processor in each path performs the first set of column and row Hartley transforms. It also obtains the power spectrum of the signal by computing the complex magnitude from the Hartley-transformed signal, as explained in Section 2.2. The subsequent processor converts the power spectrum to the log spectrum by performing a logarithmic transformation. It then performs the the second set of row and column Hartley transforms.

Finally, this processor obtains the magnitude of the cepstrum from the Hartley-transformed window. The recombination processor alternatively receives windows from the two paths. This processor performs peak detection on the cepstrum, projects the estimated disparities into the appropriate location of the disparity map, and refines the overall disparity map by a median filtering operation.

3.3. Multiple Processing Buffers. Each component of the network performs data processing in multiple processing buffers. The motivation for such an approach is to maximize the use of the Central Processing Unit (CPU) by preparing data for processing using the DMA co-processor. The processing buffers are organized to reduce the work of the CPU to mere computations without the need for considerations about data organization. The DMA co-processor transfers information into and out of the idle buffers while the CPU processes the data in the active one. These roles of "active" and "idle" rotate among all the processing buffers in a predetermined manner.

The first processor in the cepstrum computation pipeline uses four processing buffers: one to perform the Hartley transform along columns and three to perform the Hartley transform along rows and obtain the complex magnitude. The column transform buffer is used not only to perform transformations along columns but also to store those columns that are shared by neighbouring ocular dominance columns. The second processor in the pipeline uses three buffers for performing the logarithmic transformation, as well as column and row Hartley transforms on the data received from the first processor. Finally, the recombination processor uses two buffers for performing peak detection on the computed cepstrum. The result of peak detection is the value of disparity for the corresponding image patch. After peak detection, the values of the cepstrum need not be transferred to another stage and can be discarded. Therefore only two peak detection buffers are used. While peak detection is being performed on one buffer, the DMA co-processor transfers a new cepstrum into the other. There is also a processing buffer dedicated to median filtering in this processor to which the estimated disparities are written.

For the range of image and ocular dominance column sizes in our system, it is possible to use the on-chip memory of the TMS320C40 processors for processing buffers. Since the access time of on-chip memory is less than that of the off-chip memory, its use reduces the total processing time of the system. The use of on-chip memory requires satisfying the following two conditions:

$$\begin{cases} W \times H & \leq 256 \\ OPR \times (FH+1) + MDD + 2 & \leq 1024 \end{cases} \tag{9.10}$$

where W and H, respectively denote the width and height of each ocular dominance column in units of pixels. OPR, FH, and MDD represent the number of ocular dominance columns per image row, the height of the median filter, and the maximum detectable disparity, respectively. The ocular dominance dimensions provided in Chapter 5 are within this range.

3.4. Performance of the Parallel Implementation. With the above implementation, the system achieves a processing time of four seconds for a 256-pixel × 256-pixel image frame with 8-pixel × 16-pixel ocular dominance columns. Since the processing time of each ocular dominance column is fixed, the processing time of the overall frame is simply a linear function of the image size. For example, the processing time of a 64 × 64 frame would be 0.25 seconds. This is faster than the 2 second time obtained by the implementation of the algorithm in [**73**]. The latency of the system approximately equals the time required for processing one frame.

CHAPTER 10

Conclusions

In this thesis we have considered stereopsis as a component of a more complete vision system. The role of stereopsis within the overall framework of such a system is to distinguish between various objects and the background based on differences in depth. From a computational point of view, such a role for stereopsis eliminates the need for precise disparity estimation at every image point. Figure-ground separation, rather than precise distance profile determination, also translates into a stereo vision system which does not require camera calibration.

In a system which uses multiple visual cues, the final description of a scene is facilitated by all of the individual cues as well as the interaction among them. Therefore, in such a system it is important to ensure that each component, especially at the early stages, is designed to contribute to the goodness of the overall result rather than merely to the outcome of that particular stage. One of the factors determining the quality of the interaction between early stages of such a complex system is the processing time of each individual component. The low-level components can interact most efficiently if the computational time required for each of them is fixed and known a priori.

Therefore, the use of stereopsis within the context of a more elaborate vision system creates an additional computational requirement for the stereo component. To interface with the remainder of the system in a useful manner, the time required for stereo-related computations should be independent of the properties of the processed image. This thesis responds to the need for fixed processing time by posing the initial stereo disparity estimation as a local computation. This is in contrast to most traditional algorithms which use an iterative search and optimization approach.

We have employed the cepstral filtering of small patches in the stereo image pair for local estimation of disparity. The patches are referred to as ocular dominance columns and arise

in biological vision systems. We observe that choosing the dimensions of ocular dominance columns requires satisfying certain contradictory criteria. The accurate approximation of the cepstrum, the detection of the maximum disparity, and the avoidance false disparities all require large column dimensions. In contrast, a high resolution disparity map and the avoidance of multiple disparities favour small column dimensions. Assuming high image resolution, we suggest a strategy for choosing the dimensions of the ocular dominance columns based on the maximum detectable disparity.

We also indicate that inherent data correlation creates an effect similar to that caused by disparity. Based on this and a careful study of the cepstrum and its properties, we infer that local estimation can result in the detection of false disparities at some locations. In doing so, we also mention the advantages of cepstrum to autocorrelation and its data dependence. We then provide improvements to the initial disparity estimation stage based on the properties of the cepstrum. We also offer a method for refining the original disparity estimates using neighbourhood information. The overall disparity map, besides containing information about the three-dimensional structure of the scene, clearly marks the location of depth discontinuities. In this manner, it constitutes a means for figure-ground separation based on depth.

Local estimation of disparities also permits the parallel implementation of the algorithm. We use a network of TMS320C40 processors for this purpose. The implementation results in a processing speed of approximately one second for a 128-pixel $\times$ 128-pixel image. The method and the particular implementation strategy are especially appropriate for use in an active vision system. With fast and fixed running time, the stereo component can readily interface with the remaining parts of the system.

There is another important reason for coupling our approach to an active vision system. The latter facilitates the use of foveated images which have high resolution in the area of interest, the so-called fovea. This property satisfies the high resolution assumption made for choosing the dimensions of the ocular dominance columns. Also, by limiting the high resolution to the fovea rather than the whole image, foveated images reduce the computational requirements of the system. Finally, active vision systems with a focus of attention can fixate on the important regions of the scene. In such a manner the system can bring the disparities of these sections into the close-to-zero range which is most suitable for the cepstral filtering of ocular dominance columns.

1. Directions for Future Work

This thesis attempted to provide a picture of stereopsis which is consistent from several different viewpoints. In doing so, we considered the concepts of local estimation of binocular disparity and the resulting problem of false disparities, the refinement of the disparity map, and the properties of the overall depth profile. These concepts were also related to the issues of parallel computation, the use of stereopsis in a multi-cue visual system, and the functioning of active vision systems in dynamic environments. Despite the ability of the local correlation mechanism to explain neural properties, one should be careful when choosing a model for such neurons. The neurophysiological evidence examined in this thesis does imply correlation as a biologically plausible mechanism of disparity estimation. However, the model requires further neurophysiological evidence. This includes evidence for the connections required for such models and for the consistency of the input-output properties of single disparity sensitive cells within the model. Also, there are other disparity sensitive neurons, the reciprocal cells, whose properties require a more elaborate model.

Studying the mechanisms underlying disparity sensitive neurons also has potential for high computational pay-back. A mechanism such as a "filter" whose input-output properties resemble some kind of correlation measure - cepstrum for example - provides the basis for a more efficient implementation of the algorithm. Rather than computing the cepstrum using the Fourier transformation, one can then physically implement the equivalent filter and obtain much higher processing speed. Another approach to increasing the computational speed is by the use of special purpose hardware for computing the Fast Fourier Transform (FFT). Research is currently under way to obtain an approximation to the system frame rate using such an implementation.

Further support for the correlation-based disparity estimation model should also come from psychophysics. Many properties of the correlation model are known from signal processing. For example, the performance and limits of correlation in situations where the correspondence between the two images is corrupted by noise, blurring, image rotation, or other image degradation can be measured using distorted stereo pairs. Such properties can be qualitatively compared with the performance of human subjects on images which have undergone similar degradations.

Median filtering has provided a simple and effective method for refining the initial disparity map. One can develop a more elaborate strategy for the refinement stage, but perhaps at the cost of more computational time. An important characteristic of such a strategy should be its suitability for parallel implementation. Examples include strategies

similar to those which have previously been developed for the refinement of orientation of edges in an image[**89**]. A discrete relaxation algorithm [**100**] using an appropriate surface description may serve this purpose. Guidelines for such a strategy can also be sought in the interactions between the disparity sensitive neurons of biological stereo systems. Of course such an endeavour, besides needing knowledge of neuronal properties, requires accurate knowledge of the organization of the disparity sensitive neurons at higher levels of the visual cortex.

Chapter 4 pointed out that the performance of correlation and the cepstrum are dependent on the spectral properties of the processed signal. Improvements to the performance of the algorithm and its consistency may also result from image pre-processing operations. Examples of such operations are contrast enhancement or those which contribute to increasing the high frequency components of the signal. Once more, guidelines for such preprocessing operations may be obtained an examination of the early stages of the visual pathway after the retina.

Finally, it should to be noted that the use of visual memory can result in great improvements both in the quality of the scene description obtained from the visual system and in the processing speed. In a dynamic scene with an active vision system, the visual sensation at any moment may be a result of the present retinal response as well as the previous perception of the scene. For stereopsis, for example, as time passes the eyes can verge on parts of the scene, bring them into the range of small disparities, and obtain more accurate descriptions of the surface properties.

2. Concluding Remarks

Traditional stereo algorithms serve many purposes well. They can be very useful for any task which requires high accuracy but has no requirement for processing speed. An example of such an application is the construction of geographical contour maps from aerial stereo images. However, once stereopsis becomes a part of a more elaborate vision system and interacts with other visual cues to sense a dynamic environment, the requirements change. Under such circumstances, the stereo system can well afford to forfeit some accuracy in favour of properties useful to the whole system.

We induce the elegance and usefulness of the proposed system using one that already exists: the biological one. The ocular dominance columns of layer 4Cα provide a suitable structure and data representation for combining the visual information from the two eyes.

The neural connections may in turn provide the circuitry required for a correlation mechanism. A correlation disparity estimation model, such as the cepstrum, is able to explain most of the observed properties of tuned disparity selective neurons.

In discussing the biological plausibility of the method, we also provide evidence that depth perception is a result of using multiple visual cues. Little is known about the mechanisms behind the integration of such cues. However, cue integration seems a logical assumption about a visual system which provides a single perception of depth at any time. Finally, we use psychophysical findings to illustrate that the properties of the final disparity map resemble the characteristics of human stereo perception. The imprecision associated with human stereo vision, along with the importance of surface boundaries, motivates the notion that the method presented in this thesis, similar to human stereopsis, is a tool for figure-ground separation.

REFERENCES

[1] N. Ahuja and L. Abbott. Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for surface estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1007–1029, October 1993.

[2] J. Allman and S. Zucker. Cytochrome oxidase and functional coding in primate striate cortex: A hypothesis. Technical Report CIM-90-4, Center for Intelligent Machines, McGill University, Montreal, Canada, 1990.

[3] J.Y. Aloimonos, I. Weiss, and A Bandyopadhyay. Active vision. *International Journal of Computer Vision*, pages 334–356, 1987.

[4] N. Ayache and B. Faverjon. Efficient registration of stereo images by matching graph description of edge segments. *International journal of Computer Vision*, 1(2):107–131, 1987.

[5] R. Bajscy. Avtive perception vs. passive perception. In *In Proceeedings of the 3rd IEEE Workshop on Computer Vision*, pages 55–59, 1985.

[6] H.H. Baker and T.O. Binford. Depth from edge and intensity based stereo. In *In Proceeings of the 7th International Joint Conference on Artificial Intelligence*, pages 631–636, Vancouver, BC, August 1981.

[7] E. Bandari and J. Little. Detection and estimation of multiple disparities by multi-evidential correlation. Technical Report 93-38, Department of Computer Science, University of British Columbia, Vancouver, Canada, September 1993.

[8] H.B. Barlow, C. Blakemore, and J.D. Pettigrew. The neural mechanism of binocular depth discrimination. *Journal of Physiology*, 193:327–342, 1967.

[9] S.T. Barnard. A stochastic approach to stereo vision. In *In Proceedings of the 5th National Confrence on Artificial Intelligence*, pages 676–680, 1986.

[10] Stephen T. Barnard and William B. Thompson. Disparity analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):333–340, July 1980.

[11] Peter N. Belhumeur. A binocular stereo algorithm for reconstructing sloping, creased, and broken surfaces in the presence of half-occlusion. Technical Report 92-11, Harvard Robotics Laboratory, Harvard University, Cambridge, MA, October 1992.

[12] P.J. Besl. Active, optical range imaging sensors. *Machine Vision and Applications*, 1:127–152, 1988.

[13] P.O. Bishop, G.H. Henry, and C.J. Smith. Binocular interaction fields of single units in the cat striate cortex. *Journal of Physiology*, 216:39–68, 1971.

[14] P.O. Bishop and J.D. Pettigrew. Neural mechanisms of binocular vision. *Vision Research*, 26(9):1587–1600, 1986.

[15] R. Blake and H.R. Wilson. Neural models of stereoscopic vision. *Trends in Neuroscience*, 14(10):445–452, 1991.

[16] C. Blakemore. The range and scope of binocular depth discrimination in man. *Journal of Physiology*, 211:599–622, 1970.

[17] C. Blakemore, A. Fiorentini, and L. Maffei. A second neural mechanism of binocular depth discrimination. *Journal of Physiology*, 226:725–749, 1972.

[18] Bruce P. Bogert, J. R. Healy, and John W. Tukey. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In M. Rosenblatt, editor, *Proceedings of Symposium on Time Series Analysis*, pages 209–243, New York, 1963. Wiley.

[19] Marc Bolduc. A foveated sensor for robotic vision. Master's thesis, McGill University, Montreal, Quebec, Canada, 1994.

[20] R.N. Bracewell. The fast hartley transform. *Proceedings of the IEEE*, 72(8):1010–1018, August 1984.

[21] A. Brookes and K.A. Stevens. Binocular depth from surfaces versus volumes. *Journal of Experimental Psychology*, 15(3):479–484, 1989.

[22] M.J. Brooks and B.K.P. Horn. Shape and source from shading.

[23] A. Burkhalter and D.C. Van Essen. Processing ofc olor, form, and disparity information in visual areas vp and v2 of ventral extrastriate cortec in macaque monkey. *Journal of Neuroscience*, 6:2327–2351, 1986.

[24] James Cadzow. *Foundations of Digital Signal Processing*. MacMillan Publishing Company, New York, 1987.

[25] Donald G. Childers, David P. Skinner, and Robert C. Kemerait. The cepstrum: A guide to processing. *IEEE Proceedings*, 65(10):1428–1443, October 1977.

[26] A.I. Cogan, A.J. Lomakin, and A.F. Rossi. Depth in anticorrelated stereograms: Effects of spatial density and interocular delay. *Vision Research*, 33(14):1959–1975, 1993.

[27] L.K. Cormak, S.B. Stevenson, and C.M. Schor. Interocular correlation, luminance contrast, and cyclopean processing. *Vision Research*, 31:2195–2207, 1991.

[28] E.A. DeYoe and D.C. Van Essen. Segregation of efferent connections and receptive field properties in visual area v2 of the macaque. *Nature*, 317:58–61, September 1985.

[29] B.M. Dow, R. Bauer, A.Z. Snyder, and R.G. Vautin. Receptive fields and orientation shifts in the foveal striate cortex of the awake macaque monkey. In G. Edelman, W. M. Cowan, and W. E. Gall, editors, *Dynamics Aspects of of Neocortical Function*, chapter 2, pages 41–65. Wiley, New York, 1984.

[30] John Ens and Ze-Nian Li. Real time motion stereo. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 130–135, New York City, NY, June 1993. IEEE Computer Society Press.

[31] M. Fendick and G. Westheimer. Effects of practice and the separation of targets on foveal and peripheral stereoacuity. *Vision Research*, 23(2):145–150, 1983.

[32] D. Ferster. A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex. *Journal of Physiology*, 311:623–655, 1981.

[33] D.J. Fleet and A.D. Jepson. Stability of phase information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12):1253–1268, 1993.

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

[35] Pascal Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine Vision and Applications*, 6(1):35–49, 1993.

[36] D.B. Gennery. A stereo vision for autonomous vehicles. In *In Proceedings of the 5th International Joint Confrence on Artificial Inelligence*, pages 576–582, Cambridge, MA, 1977.

[37] D.B. Gennery. Object detection and measurement using stereo vision. In *In Proceedings of the ARPA Image Understanding Workshop*, pages 161–167, Colege Park, MD, 1980.

[38] B. Gillam, D. Chambers, and T. Russo. Postfusional latency in stereoscopic slant perception and the primitives of stereopsis. *Journal of Experimental Psychology: Human Perception and Perofrmance*, 14(2):163–175, 1988.

[39] B. Gillam, T.Flagg, and D. Finlay. Evidence for disparity change as the primary stimulus for stereoscopic processing. *Perception and Psychophysics*, 36(6):559–564, 1984.

[40] W.E.L. Grimson. A computer implementation of a theory of human stereo vision. In *Philosophical Transactions of the Royal Society of London*, pages 217–253, 1981.

[41] W.E.L. Grimson. Computational experiments with a feature-based stereo algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):121–126, January 1985.

[42] W.E.L. Grimson. Why stereo vision is not always about 3d reconstruction. Technical Report A.I. Memo No. 1435, Massachusetts Institute of Technology, Cambridge, MA, 1993.

[43] M. J. Hannah. *Computer Matching of Areas in Stereo Imagery*. PhD thesis, Stanford University, Stanford, CA, 1974.

[44] M. J. Hannah. Bootstrap stereo. In *In Proceedings of ARPA Image Understanding Workshop*, pages 201–208, College Park, MD, April 1980.

[45] M. J. Hannah. Sri's baseline stereo system. In *In Proceedings of DARPA Image Understanding Workshop*, pages 149–155, Miami Beach, FL, December 1985.

[46] J.M. Harris and A.J. Parker. Objective evaluation of human and computational stereoscopic visual systems. *Vision Research*, 34(20):2773–2785, 1994.

[47] T.S. Haung, G.J. Yang, and G.Y. Tang. A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):13–18, February 1979.

[48] Simon Haykin. *An Introduction to Analog and Digital Communications*. John Wiley and Sons, New York, 1989.

[49] B.K.P. Horn. Understanding image intensities. *Artificial Intelligence*, 8:201–231, 1977.

[50] D.H. Hubel and M.S. Livingstone. Segregation of form, color, and stereopsis in primate area 18. *Journal of Neuroscience*, 7(11):3378–3415, 1987.

[51] D.H. Hubel and T.N. Wiesel. Functional architecture of macaque monkey visual cortex. In *In Proceedings of the Royal Society of London*, volume 198, pages 1–59, July 1977.

[52] R. A. Hummel and S. W. Zucker. On the foundations of relaxation labelling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3):267–287, May 1983.

[53] M. R. M. Jenkin and A. D. Jepson. Recovering local surface structure through local phase difference measurements. *CVGIP*. Accepted for Publication.

[54] David J. Jones. *Computational Models of Binocular Vision*. PhD thesis, Stanford University, 1991.

[55] David J. Jones and David G. Lamb. Analyzing the visual echo: Passive 3-d imaging with a multiple aperture camera. Technical Report CIM-93-3, Center For Intelligent Machines, McGill University, Montreal, Canada, February 1993.

[56] David J. Jones and Jitendra Malik. A computational framework for determining stereo correspondence from a set of linear spatial filters. In *Proceedings of the European Conference on Computer Vision*, pages 395–410, Geneva, Italy, May 1992.

[57] D.E. Joshua and P.O. Bishop. Binocular single vision and depth discrimination. receptive field disparities for central and peripheral vision and binocular interaction on peripheral single units in cat striate cortex. *Experimental Brain Research*, 10:389–416, 1970.

[58] B. Julesz. Stereoscopic vision. *Vision Research*, 26(9):1601–1612, 1986.

[59] M. Kass. Computing visual correspondence. In *In Proc. DARPA Image Understanding Workshop*, pages 54–60, Arlington, VA, June 1983.

[60] A.E. Kertesz and D.R. Hampton. Fusional response to extrafoveal stimulation. *Investigative Ophthalmology and Visual Science*, 21:600–605, 1981.

[61] Y.C. Kim and J.K. Aggarwal. Positioning 3-d objects using stereo images. *IEEE Journal of Robotics and Automation*, RA-3(4):361–373, August 1987.

[62] E. Kreyszig. *Advanced Engineering Mathematics*. John Wiley and Sons, New York, NY, sixth edition, 1988.

[63] K. Langley, T.J. Atherton, R.G. Wilson, and M.H.E. Larcombe. Vertical and horizontal disparities from phase. In *In Proceeedings of the 1st European Conference on Computer Vision*, pages 315–325, Antibes, 1990.

[64] M. D. Levine. *Vision in Man and Machine*. McGraw Hill, Toronto, 1985.

[65] Martin D. Levine, Douglas A. O'Handley, and Gary M. Yagi. Computer determination of depth maps. *Computer Graphics and Image Processing*, 2(2):131–150, October 1973.

[66] Kai-Oliver Ludwig and Bernd Neumann Heiko Neumann. Local stereoscopic depth estimation using ocular stripe maps. In *Proceedings of the European Conference on Computer Vision*, pages 373–377, Geneva, Italy, May 1992.

[67] S. B. Marapane and M. M. Trivedi. Region-based stereo analysis for robotics applications. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1447–1464, November–December 1989.

[68] D. Marr and T. Poggio. A computational theory of human stereo vision. *Science*, 194:283–287, 1976.

[69] D. Marr and T. Poggio. A theory of human stereo vision. MIT A.I. Memo No. 451, November 1977.

[70] D. Marr and T. Poggio. A computational theory of human stereo vision. In *In Proceedings of the Royal Society of London*, volume B204, pages 301–328, 1979.

[71] David Marr. *Vision*, chapter 3, pages 111–159. W.H. Freeman and Company, New York, 1982.

[72] R. Maske, S. Yamane, and P.O. Bishop. Binocular simple cells for local stereopsis: Comparison of receptive field organizations for the two eyes. *Vision Research*, 24(12):1921–1929, 1984.

[73] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal of Computer Vision*, 8(1):71–91, 1992.

[74] J. E. W. Mayhew and J. P. Frisby. Psychological and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, 17:349–385, 1981.

[75] J.E.W. Mayhew and H.C. Longuet-Higgins. A computational model of binocular depth perception. *Nature*, 297:376–378, 1982.

[76] S.P. McKee, D.M. Levi, and S.F. Bowne. The imprecision of stereopsis. *Vision Research*, 30(11):1763–1769, 1990.

[77] D.E. Mitchell. A review of the concept of "panum's fusional area". *American Journal of Optometry and Archives of American Academy of Optometry*, pages 387–401, 1966.

[78] G.J. Mitchison and G. Westheimer. The perception of depth in simple figures. *Vision Research*, 24(9):1063–1073, 1984.

[79] H.P. Moravec. Towards automatic visual obstacle avoidance. In *In Proceedings of the 5th International Joint Confrence on Srtificial Intelligence*, page 584, Cambridge, MA, 1977.

[80] G. Medioni na dR. Nevatia. Segment-based stereo matching. *Computer Vision, Graphics, and Image Processing*, 31:2–18, 1985.

[81] T. Nikara, P.O. Bishop, and J.D. Pettigrew. Analysis of retinal correspondence by studying receptive fields of binocular single units in cat striate cortex. *Experimental Brain Research*, 6:353–372, 1968.

[82] H. K. Nishihara. Practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545, September-October 1984.

[83] K.N. Ogle. Disparity limits of stereopsis. *Archives of Ophthalmology*, 48(50–60), 1952.

[84] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 63–69, Hawaii, June 1991. IEEE Computer Society Press.

[85] T. J. Olson and D. J. Coombs. Real time vergence control for binocular robots. *International Journal of Computer Vision*, 7(1):67–89, November 1991.

[86] T. J. Olson and R. D. Potter. Real time vergence control. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 404–409, San Diego, June 1989. IEEE Computer Society Press.

[87] Alan V. Oppenheim and Ronald W. Schafer. *Discrete Time Signal Processing*, chapter 12, pages 768–834. Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[88] Alan V. Oppenheim and Ronald W. Shafer. *Discrete Time Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[89] P. Parent and S. W. Zucker. Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):823–839, August 1989.

[90] A.P. Pentland. Local shading analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):170–187, March 1984.

[91] G. F. Poggio, F. Gonzalez, and F. Krause. Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity. *Journal of Neuroscience*, 8(12):4531–4550, 1988.

[92] G. F Poggio and T Poggio. The analysis of stereopsis. *Annual Review of Neuroscience*, 7:379–412, 1984.

[93] G. F. Poggio and W. H. Talbot. Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey. *Journal of Physiology*, 315:469–492, 1981.

[94] G.F. Poggio. Processing of stereoscopic information in primate visual cortex. In G. Edelman, W. M. Cowan, and W. E. Gall, editors, *Dynamics Aspects of of Neocortical Function*, chapter 20, pages 613–635. Wiley, New York, 1984.

[95] G.F. Poggio and B. Fischer. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology*, 40(6):1392–1405, November 1977.

[96] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[97] S.C. Rawlings and T. Shipley. Stereoscopic acuity and horizontal angular distance from fixation. *Journal of the Optical Society of America*, 59(8):991–993, August 1969.

[98] D. Regan, J.P. Frisby, G.F. Poggio, C.M. Schor, and C.W. Tyler. The perception of stereodepth and stereomotion. In L. Spillmann and J.S. Werner, editors, *Visual Perception: The Neurophysiological Foundations*, chapter 12, pages 317–347. Academic Press, 1990.

[99] B.J. Rogers and M.F. Bradshaw. Vertical disparities, differential perspective and binocular stereopsis. *Nature*, 361:253–255, January 1993.

[100] A. Rosenfeld, R. A. Hummel, and S. W. Zucker. Scene labelling by relaxation processes. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6:420–433, 1976.

[101] Bill Ross. A practical stereo vision system. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 130–135, New York City, NY, June 1993.

[102] T.D. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418, 1988.

[103] P.H. Schiller. Central connections of the retinal on and off pathways. *Nature*, 297:580–582, June 1982.

[104] C. Schor, T. Heckmann, and C.W. Tyler. Binocular fusion limits are independent of contrast, luminance gradient and component phases. *Vision Research*, 29(7):821–835, 1989.

[105] C. Schor, M. Wesson, and K.M. Robertson. Combined effects of spatial frequency and retinal eccentricity upon fixation disparity. *American Journal of Optometry and Physiological Optics*, 63(8):619–626, 1986.

[106] C.M. Schor and D.R. Badcock. A comparison of stereo and vernier acuity within spatial channels as a function of distance from fixation. *Vision Research*, 25(8):1113–1119, 1985.

[107] C.M. Schor and C.W. Tyler. Spatio-temporal properties of panum's fusional area. *Vision Research*, 21:683–692, 1980.

[108] E.L. Schwartz. Columnar architecture and computational anatomy in primate visual cortex: Segmentation and feature extraction via spatial frequency coded difference mapping. *Biological Cybernetics*, 42:157–168, 1982.

[109] Eric L. Schwartz and Yehezkel Yeshurun. Cortical hypercolumn size determines stereo fusion limits. Submitted to the Journal of Computational Neuroscience, November 1991.

[110] Gal Sela. Real-time attention for robotic vision. Master's thesis, McGill University, Montreal, Quebec, Canada, 1995.

[111] Samir Shah and Martin D. Levine. Visual information processing in primate cone pathways: Part I, a model. *IEEE Transactions on Systems, Man and Cybernetics*, 1995. Accepted for Publication.

[112] Samir Shah and Martin D. Levine. Visual information processing in primate cone pathways: Part II, experiments. *IEEE Transactions on Systems, Man and Cybernetics*, 1995. Accepted for Publication.

[113] T. Shipley and M. Popp. Stereoscopic acuity and retinal eccentricity. *Ophthalmic Research*, 3:252–255, 1972.

[114] M.C. Steckner and D.J. Drost. Fast cepstrum analysis using the hartley transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(8):1300–1302, August 1989.

[115] K.A. Stevens. Constructing the perception of surfaces from multiple cues. *Mind and Language*, 5(4):253–266, 1990.

[116] K.A. Stevens and A. Brookes. Depth reconstruction in stereopsis. In *Proceedings of the International Conference on Computer Vision*, pages 682–686, London, England, June 1987.

[117] K.A. Stevens and A. Brookes. Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research*, 28(3):371–386, 1988.

[118] K.A. Stevens, M. Lees, and A. Brookes. Combining binocular and monocular curvature feature. *Perception*, 20:425–440, 1991.

[119] S.B. Stevenson, L.K. Cormack, and C.M. Schor. The effect of stimulus contrast and interocular correlation on disparity vergence. *Vision Research*, 34(3):383–396, 1994.

[120] R.B. Tootel, M.S. Silverman, and R.L. DeValois. Spatial frequency columns in primary visual cortex. *Science*, 214:813–815, November 1981.

[121] C.W. Tyler. Spatial organization of binocular disparity. *Vision Research*, 15:583–590, 1975.

[122] J. Weng. A theory of image matching. In *In Proceedings of the 3rd Internatinal Conference in Computer Vision*, pages 200–209, Osaka, Japan, 1990.

[123] R.H. Williams. *Electrical Engineering Probability*. West Publishing Company, St. Paul, MN, 1991.

[124] H.R. Wilson, D. Levi, L. Maffei, J. Rovamo, and R. DeValois. The perception of form: Retina to striate cortex. In L. Spillmann and J.S. Werner, editors, *Visual Perception: The Neurophysiological Foundations*, chapter 10, pages 231–272. Academic Press, 1990.

[125] Yehezkel Yeshurun and Eric L. Schwartz. Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):759–767, July 1989.

[126] S.W. Zucker, A. Dobbins, and L. Iverson. Two stages of curve detection suggest two styles of visual computation. *Neural Computation*, 1:68–81, 1989.

[127] S.W. Zucker and R.A. Hummel. Receptive fields and the representation of the visual information. *Human Neurobiology*, 5:121–128, 1986.

APPENDIX A

Autocorrelation and Power Spectrum

Equation 4.1 defines the autocorrelation function of a signal $g(t)$ as the convolution of the signal with its reflected version:

$$R_g(\tau) = g(t) * g(-t) \tag{A.1}$$

Using the convolution property of the Fourier transform

$$\mathcal{F}[R_g(\tau)] = \mathcal{F}[g(t)] \cdot \mathcal{F}[g(-t)] \tag{A.2}$$

where $\mathcal{F}$ denotes Fourier transformation. Using the definition of Fourier transform

$$\mathcal{F}[g(-t)] = \int_{-\infty}^{+\infty} g(-t)e^{-j2\pi ft}dt \tag{A.3}$$

Substituting t by $-t$ gives

$$\mathcal{F}[g(-t)] = -\int_{+\infty}^{-\infty} g(t)e^{j2\pi ft}dt = \int_{-\infty}^{+\infty} g(t)e^{j2\pi ft}dt \tag{A.4}$$

Since $g(t)$ is a real signal, its complex conjugate, $g^*(t)$, equals $g(t)$. Therefore

$$\mathcal{F}[g(-t)] = \left(\int_{-\infty}^{+\infty} g(t)e^{-j2\pi ft}dt\right)^* \tag{A.5}$$

But the right hand side of Equation A.5 is simply the complex conjugate of the Fourier transform of $g(t)$. Hence

$$\mathcal{F}[g(-t)] = (G(f))^* \tag{A.6}$$

Substituting into Equation A.2 gives

$$\mathcal{F}[R_g(\tau)] = G(f) \cdot (G(f))^* = \mid G(f) \mid^2 \tag{A.7}$$

which proves Equation 4.2.

APPENDIX B

The Cepstrum of a Signal Containing an Echo

An echo is defined as a shifted and possibly scaled version of the original signal. Therefore, it can be described using convolution with a shifted delta function. Also using Equation 4.7, one can represent an original signal $s(x)$ with an echo added to it as

(B.1) $$g(x) = s(x) * (\delta(x) + a\delta(x - D))$$

or equivalently, in Equation 4.7

(B.2) $$f(x) = \delta(x) + a\delta(x - D)$$

where D is the shift which generates the echo and a is a scaling factor. Taking the Fourier transform of both sides and using the convolution and time domain shift properties of the Fourier transform gives

(B.3) $$| G(f) | = | S(f) | . | \left(1 + ae^{-j2\pi fD}\right) |$$

Taking the logarithm yields

(B.4) $$\log | G(f) | = \log | S(f) | + \log \left(1 + ae^{-j2\pi fD}\right)$$

Using the expansion

(B.5) $$\log(1 + z) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{z^m}{m}$$

which is valid for $| z | \leq 1$ and $z \neq -1$, in Equation B.4 gives

$$\log \mid G(f) \mid = \log \mid S(f) \mid + \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\left(ae^{-j2\pi fD}\right)^m}{m} \tag{B.6}$$

The last term in Equation B.6 appears as a set of ripples in the log-spectrum. The frequency of these ripples in the frequency domain, or their quefrency, is proportional to the shift in the echo D.

Evaluating the inverse Fourier transform of Equation B.6 gives

$$g_{cep}(x) = \int_{-\infty}^{\infty} \left(\log \mid S(f) \mid + \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\left(ae^{-j2\pi fD}\right)^m}{m} \right) e^{j\pi fx} df \tag{B.7}$$

From the linearity of integration we have

$$g_{cep}(x) = \int_{-\infty}^{\infty} \log \mid S(f) \mid e^{j\pi fx} df + \int_{-\infty}^{\infty} e^{j\pi fx} \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\left(ae^{-j2\pi fD}\right)^m}{m} df \tag{B.8}$$

which by definition of the power cepstrum and the properties of the inverse Fourier transform is equivalent to

$$g_{cep}(x) = s_{cep}(x) + \sum_{m=1}^{\infty} (-1)^{m+1} a^m \frac{\delta(x - mD)}{m} \tag{B.9}$$

which contains impulses at multiples of echo delay.

APPENDIX C

Cepstral Peak Magnitude Ratios

Assume that a resultant signal contains an uncorrelated original and an echo with disparity d. Further assume that the ratio of the maxima of the autocorrelation of the resultant signal when $d \neq 0$ and $d = 0$ is equal to a, $0 < a < 1$. The autocorrelation function of the resultant signal can then be represented by

$$R(\tau) = \delta(\tau) + a\delta(\tau + d) + a\delta(\tau - d) \tag{C.1}$$

Of course, the unshifted unit impulse in Equation C.1 is due to the fact that every signal, whether it contains and echo or not, is correlated with itself at a shift of zero. The shifted and scaled impulses are generated by the presence of the echo.

As described by Equation C.2, the cepstrum of a signal can be obtained using the analogy to a power series. Each term of the series contains multiple convolutions, rather than multiplications, of the autocorrelation of the original signal. The ratio of the cepstral peak magnitude at shift d to that of shift zero then provides the drop in the cepstral peak magnitude from disparity zero to d.

$$g_{cep}(x) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{(R_g(\tau))^{m*}}{m} \tag{C.2}$$

Convolving the autocorrelation function of Equation C.1 with itself results in a recursive equation. Denoting the $(m-1)$th convolution of $R(\tau)$ with itself by f^{m*}, this relationship can be described as

$$(C.3)\quad f^{m*}(kd) = \begin{cases} 0 & |\,k\,| > m \\ af^{m-1}\left((k-1)d\right) & k = \pm m \\ af^{m-1}\left((k-1)d\right) + f^{m-1}(kd) & k = \pm(m-1) \\ af^{m-1}\left((k-1)d\right) + f^{m-1}(kd) + af^{m-1}\left((k+1)d\right) & |\,k\,| < (m-1) \end{cases}$$

The magnitudes of the cepstral peaks at shifts d and zero are $f^{m*}(d)$ and $f^{m*}(0)$ respectively. Due to the "diffusing nature" of the above equation, its solution is rather tedious. Looking at the original problem, we notice that obtaining the result of multiple self-convolutions readily lends itself to frequency domain analysis.

One can consider the autocorrelation function of Equation C.1 as the distribution of a random variable (the function can easily be normalized so that its integral equals unity). Then, $m-1$ convolutions of this function with itself represents the distribution of a new random variable. This new random variable is obtained by adding m random variables of the former distribution together. The characteristic function of the resultant random variable is the product of those of the initial ones. Noting that the characteristic function of a random variable equals the Fourier transform of its distribution [**123**], we can write

$$(C.4)\qquad \mathcal{F}[f^{m*}(x)] = (\mathcal{F}[R(\tau)])^m$$

Equation C.4 is of course a statement of the convolution-multiplication duality of the Fourier transformation. Defining the δ-function as a distribution ensures that it receives the extra care required in frequency domain operations.

Using the shift property of the Fourier transform, the transform of the autocorrelation function of Equation C.1 can be written as

$$(C.5)\qquad \mathcal{F}[R(\tau)] = 1 + ae^{-j\pi fd} + ae^{j\pi fd} = 1 + 2a\cos(\pi fd)$$

Substituting this value into Equation C.4 gives

$$\begin{aligned} \mathcal{F}[f^{m*}(x)] &= (1 + 2a\cos(\pi fd))^m \\ (C.6) &= 1 + \binom{m}{1} 2a\cos(\pi fd) + \cdots + \binom{m}{k} (2a)^k \cos^k(\pi fd) + \cdots + (2a)^m cos^m(\pi fd) \end{aligned}$$

Inverse Fourier transforming the expression of Equation C.6 results in a series of δ-functions at multiples of shift d. The magnitude of the δ-functions at any shift contributes

to the value of cepstrum at that shift through the relationship of Equation C.2. Therefore, the values of cepstrum at shifts zero and d, depend on the constant and $\cos(\pi f d)$ terms of Equation C.6 respectively. In obtaining the coefficients of these two from Equation C.6, it is important to note that any term $\cos^k(\pi f d)$ contains either the zeroth or the first harmonic of $\cos(\pi f d)$ depending on whether k is odd or even. The coefficients of the these harmonics can be obtained in the following manner.

$$
\begin{aligned}
cos^k(\pi f d) &= \frac{1}{2^k}\left(e^{j\pi f d} + e^{-j\pi f d}\right)^k \\
&= \frac{1}{2^k}\left(e^{jk\pi f d} + \binom{k}{1} e^{j(k-1)\pi f d} e^{-j\pi f d} + \cdots + \binom{k}{n} e^{j(k-n)\pi f d} e^{-jn\pi f d} + \right. \\
&\quad \left. \cdots + \binom{k}{k-1} e^{j\pi f d} e^{-j(k-1)\pi f d} + e^{-jk\pi f d}\right) \\
&= \frac{1}{2^k}\left(e^{jk\pi f d} + \binom{k}{1} e^{j(k-2)\pi f d} + \cdots + \binom{k}{n} e^{j(k-2n)\pi f d} + \right. \\
&\quad \left. \cdots + \binom{k}{k-1} e^{-j(k-2)\pi f d} + e^{-jk\pi f d}\right)
\end{aligned}
\tag{C.7}
$$

Equation C.7 contains the first harmonic of $\cos(\pi f d)$ if k is odd and the zeroth if it is even. For odd k, the middle two terms of Equation C.7 are

$$
\begin{aligned}
&\frac{1}{2^k}\left(\cdots + \binom{k}{\frac{k-1}{2}} e^{j(k-\frac{k-1}{2})\pi f d} e^{-j(\frac{k-1}{2})\pi f d} + \binom{k}{\frac{k+1}{2}} e^{j(k-\frac{k+1}{2})\pi f d} e^{-j(\frac{k+1}{2})\pi f d} + \cdots\right) \\
&\qquad = \frac{1}{2^k}\left(\cdots + \binom{k}{\frac{k-1}{2}} e^{j\pi f d} + \binom{k}{\frac{k+1}{2}} e^{-j\pi f d} + \cdots\right)
\end{aligned}
\tag{C.8}
$$

But for odd k

$$
\binom{k}{\frac{k-1}{2}} = \binom{k}{\frac{k+1}{2}} = \frac{k!}{(\frac{k-1}{2})!(\frac{k+1}{2})!}
\tag{C.9}
$$

Replacing this value in Equation C.8 gives

$$
\begin{aligned}
cos^k(\pi f d) &= \frac{1}{2^k}\left(\cdots + \frac{k!}{\left(\frac{k-1}{2}\right)!\left(\frac{k+1}{2}\right)!}\left(e^{j\pi f d} + e^{-j\pi f d}\right) + \cdots\right) \\
&= \frac{1}{2^k}\left(\cdots + \frac{k!}{\left(\frac{k-1}{2}\right)!\left(\frac{k+1}{2}\right)!} 2cos(\pi f d) + \cdots\right) \\
&= \frac{1}{2^{k-1}}\left(\cdots + \frac{k!}{\left(\frac{k-1}{2}\right)!\left(\frac{k+1}{2}\right)!} cos(\pi f d) + \cdots\right) \qquad \text{(C.10)}
\end{aligned}
$$

For even k, with similar steps one can illustrate that $\cos^k(\pi f d)$ contains the zeroth (DC) harmonic of $\cos(\pi f d)$ in the following form

$$
cos^k(\pi f d) = \frac{1}{2^k}\left(\cdots + \frac{k!}{\left(\frac{k}{2}\right)!\left(\frac{k}{2}\right)!} cos(\pi f d) + \cdots\right) \qquad \text{(C.11)}
$$

Replacing the value of Equation C.10 into Equation C.6 gives

$$
\begin{aligned}
&\binom{m}{1} 2a\cos(\pi f d) + \textstyle\sum_{k=3,k\ odd}^{2*\lfloor\frac{m+1}{2}\rfloor} \binom{m}{k} \frac{(2a)^k}{2^{k-1}} \binom{k}{\frac{k-1}{2}} \cos(\pi f d) \\
= \;& 2\cos(\pi f d)\left(\binom{m}{1} a + \textstyle\sum_{i=1}^{\lfloor\frac{m-1}{2}\rfloor} a^{2i+1} \frac{m!}{(2i+1)!(m-2i-1)!} \frac{(2i+1)!}{i!(2i+1-i)!}\right) \\
= \;& 2\cos(\pi f d)\left(\binom{m}{1} a + \textstyle\sum_{i=1}^{\lfloor\frac{m-1}{2}\rfloor} a^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!}\right) \\
= \;& (e^{j\pi f d} + e^{-j\pi f d}) \textstyle\sum_{i=0}^{\lfloor\frac{m-1}{2}\rfloor} a^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!} \qquad \text{(C.12)}
\end{aligned}
$$

as the magnitude of the first harmonic of $\cos^k(\pi f d)$ in $\mathcal{F}[f^{m*}(x)]$. Replacing Equation C.11 into Equation C.6 gives

$$
\begin{aligned}
& 1 + \sum_{k=0,k\ even}^{2*\lfloor \frac{m}{2} \rfloor} \binom{m}{k} \frac{(2a)^k}{2^k} \binom{k}{\frac{k}{2}} \\
= \ & 1 + \sum_{i=1}^{\lfloor \frac{m}{2} \rfloor} a^{2i} \frac{m!}{(2i)!(m-2i)!} \frac{(2i)!}{i!(2i-i)!} \\
= \ & 1 + \sum_{i=1}^{\lfloor \frac{m}{2} \rfloor} a^{2i} \frac{m!}{(m-2i)!i!i!} \\
= \ & \sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} a^{2i} \frac{m!}{(m-2i)!i!i!} \qquad \text{(C.13)}
\end{aligned}
$$

as the magnitude of the DC value in $\mathcal{F}[f^{m*}(x)]$.

Taking the inverse Fourier Transform of $\mathcal{F}[f^{m*}(x)]$ gives the magnitudes of the impulses at shifts zero and d in $f^{m*}(x)$ as

$$Peak_{ceps}(d) = \sum_{i=0}^{\lfloor \frac{m-1}{2} \rfloor} a^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!} \qquad \text{(C.14)}$$

$$Peak_{ceps}(0) = \sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} a^{2i} \frac{m!}{(m-2i)!i!i!} \qquad \text{(C.15)}$$

Normalizing the peak at shift d by that at zero and placing the normalized peak in Equation 4.17 gives

$$\frac{Peak_{ceps}(d)}{Peak_{ceps}(0)} = \begin{cases} 1 & d = 0 \\ \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{\sum_{i=0}^{\lfloor \frac{m-1}{2} \rfloor} (a)^{2i+1} \frac{m!}{(m-2i-1)!i!(i+1)!}}{\sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} (a)^{2i} \frac{m!}{(m-2i)!i!i!}} & d \neq 0 \end{cases} \qquad \text{(C.16)}$$

Replacing a by the appropriate autocorrelation ratio proves Equations 4.18 and 4.19.

APPENDIX D

Hartley Transform and Magnitude Spectrum

The power spectrum of a two-dimensional signal, $f(m,n)$ is obtained from its Fourier transform as follows:

$$\begin{aligned} \mid F(u,v) \mid^2 &= (Re[F(u,v)])^2 + (Im[F(u,v)])^2 \\ & \qquad\qquad u = 0,1,\cdots,M-1; v = 0,1,\cdots,N-1 \end{aligned} \tag{D.1}$$

Equivalently,

$$\begin{aligned} \mid F(u,v) \mid^2 &= \left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\cos\left(2\pi\left(\tfrac{um}{M}+\tfrac{vn}{N}\right)\right)\right)^2 \\ &\quad + \left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\sin\left(2\pi\left(\tfrac{um}{M}+\tfrac{vn}{N}\right)\right)\right)^2 \\ &\qquad u = 0,1,\cdots,M-1; v = 0,1,\cdots,N-1 \end{aligned} \tag{D.2}$$

From the definition of the Hartley transform, one can obtain the following four relationships for $u = 0,1,\cdots,M-1; v = 0,1,\cdots,N-1$:

$$\begin{aligned} H(u,v) &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(\tfrac{2\pi um}{M}\right)+\sin\left(\tfrac{2\pi um}{M}\right)\right)\left(\cos\left(\tfrac{2\pi vn}{N}\right)+\sin\left(\tfrac{2\pi vn}{N}\right)\right) \\ &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(2\pi\left(\tfrac{um}{M}-\tfrac{vn}{N}\right)\right)+\sin\left(2\pi\left(\tfrac{um}{M}+\tfrac{vn}{N}\right)\right)\right) \end{aligned} \tag{D.3}$$

$$
\begin{aligned}
H(M-u,v) &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) \\
&\quad \left(\cos\left(\frac{2\pi(M-u)m}{M}\right)+\sin\left(\frac{2\pi(M-u)m}{M}\right)\right)\left(\cos\left(\frac{2\pi vn}{N}\right)+\sin\left(\frac{2\pi vn}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(\frac{2\pi um}{M}\right)-\sin\left(\frac{2\pi um}{M}\right)\right)\left(\cos\left(\frac{2\pi vn}{N}\right)+\sin\left(\frac{2\pi vn}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(2\pi\left(\frac{um}{M}-\frac{vn}{N}\right)\right)+\sin\left(2\pi\left(\frac{vn}{N}-\frac{um}{M}\right)\right)\right)
\end{aligned}
$$

(D.4)

$$
\begin{aligned}
H(u,N-v) &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) \\
&\quad \left(\cos\left(\frac{2\pi um}{M}\right)+\sin\left(\frac{2\pi um}{M}\right)\right)\left(\cos\left(\frac{2\pi(N-v)n}{N}\right)+\sin\left(\frac{2\pi(N-v)n}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(\frac{2\pi um}{M}\right)+\sin\left(\frac{2\pi um}{M}\right)\right)\left(\cos\left(\frac{2\pi vn}{N}\right)-\sin\left(\frac{2\pi vn}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\left(\cos\left(2\pi\left(\frac{um}{M}-\frac{vn}{N}\right)\right)+\sin\left(2\pi\left(\frac{um}{M}-\frac{vn}{N}\right)\right)\right)
\end{aligned}
$$

(D.5)

$$
\begin{aligned}
H(M-u,N-v) &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) \\
&\quad \left(\cos\left(\frac{2\pi(M-u)m}{M}\right)+\sin\left(\frac{2\pi(M-u)m}{M}\right)\right) \\
&\quad \left(\cos\left(\frac{2\pi(N-v)n}{N}\right)+\sin\left(\frac{2\pi(N-v)n}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0} N-1 f(m,n)\left(\cos\left(\frac{2\pi um}{M}\right)-\sin\left(\frac{2\pi um}{M}\right)\right) \\
&\quad \left(\cos\left(\frac{2\pi vn}{N}\right)-\sin\left(\frac{2\pi vn}{N}\right)\right) \\
&= \sum_{m=0}^{M-1}\sum_{n=0} N-1 f(m,n)\left(\cos\left(2\pi\left(\frac{um}{M}-\frac{vn}{N}\right)\right)-\sin\left(2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\right)\right)
\end{aligned}
$$

(D.6)

The values of $\cos\left(2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\right)$ and $\sin\left(2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\right)$ can be obtained from Equation D.3 to D.6 as follows:

$$f(m,n)\cos\left(2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\right) = \frac{1}{2}\left(H(M-u,v)+H(u,N-v)\right) \tag{D.7}$$

$$f(m,n)\sin\left(2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)\right) = \frac{1}{2}\left(H(u,v)-H(M-u,N-v)\right) \tag{D.8}$$

Note that $H(M,v)=H(0,v)$ and $H(u,N)=H(u,0)$. Substituting these values in Equation D.2 gives

(D.9)

$$\begin{cases} |F(0,0)|^2 = (H(0,0))^2 \\ |F(u,0)|^2 = \frac{1}{4}\left((H(u,0)+H(M-u,0))^2+(H(u,0)-H(M-u,0))^2\right) \\ \qquad\quad = \frac{1}{2}\left(H^2(u,0)+H^2(M-u,0)\right) \qquad u=1,\cdots,M-1 \\ |F(0,v)|^2 = \frac{1}{4}\left((H(0,v)+H(0,N-v))^2+(H(0,v)-H(0,N-v))^2\right) \\ \qquad\quad = \frac{1}{2}\left(H^2(0,v)+H^2(0,N-v)\right) \qquad v=1,\cdots,N-1 \\ |F(u,v)|^2 = \frac{1}{4}\left((H(M-u,v)+H(u,N-v))^2+(H(u,v)-H(M-u,N-v))^2\right) \\ \qquad\qquad\qquad u=1,\cdots,M-1; v=1,\cdots,N-1 \end{cases}$$

which proves Equation 9.9.