Passive Monocular Range Imaging with a Multiple Aperture Camera

David G. Lamb

Department of Electrical Engineering

McGill University

Montréal

August, 1994

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Engineering

ISBN 0-315-99971-3

Canada

Abstract

When the iris of a conventional camera is replaced by a mask with multiple apertures, a composite image is formed. Unlike binocular stereopsis, the views from each aperture are superimposed, so that conventional methods in stereo vision do not apply. Still, the local displacement between corresponding points in these views is related to their distance from the camera. This depth cue provides the basis for a new paradigm in passive range sensing — *monocular stereopsis*. This thesis presents a technique for computing a dense range image from one composite image acquired with a multiple aperture camera. The formation of the composite image is modelled as an echo process, where the depth of a point in the scene is directly related to the spatial delay of its visual echo. Cepstral analysis is the method used to detect this echo. A model of the composite image cepstrum allows measurement of monocular disparity to subpixel precision, as well as an estimate of its associated error distribution. This data, computed over a dense grid, is used to generate a piecewise planar representation of surfaces in the scene, based on a maximum likelihood criterion. Borrowing techniques from visual psychophysics, the spatial resolution of this result is evaluated in terms of an intelligent agent making decisions about its environment. This new range imaging technique is successfully applied to real-world scenes to demonstrate its potential for mobile robot navigation and obstacle avoidance.

Résumé

Lorsque le diaphragme d'une caméra conventionnelle est remplacé par un masque avec plusieurs ouvertures, une image composite se forme. Contrairement à la stéréoscopie binoculaire, les vues des différentes ouvertures se superposent. Les méthodes conventionnelles en vision stéréoscopique ne s'appliquent donc pas. Néanmoins, l'écart dans l'image des différentes vues d'un point est relié à sa distance de la caméra. Ce signal de profondeur fournit la base d'un nouveau paradigme en télémétrie passive — la stéréoscopie monoculaire. Cette thèse présente une technique pour calculer une carte de profondeur dense à partir d'une image composite obtenue par une camera à ouvertures multiples. La formation de l'image composite est modèlisée comme un procédé d'écho pour lequel la profondeur d'un point dans la scène est directement relié au délais spatial de l'écho visuel. L'analyse "cepstrale" est la méthode utilisée pour détecter cet echo. Un modèle du "cepstrum" de l'image composite permet la mesure de la disparité monoculaire avec une précision plus petite qu'un pixel, ainsi qu'un estimé de la distribution de l'incertitude. Cette donnée, calculée sur une grille serrée, est utilisée pour générer une représentation de la scène en se basant sur un critère de probabilité maximale. Nous empruntons une technique utilisée à la psycho-physique visuelle pour évaluer la résolution spatiale de ces résultats en terme d'un agent intelligent prenant des décisions concernant son environnement. Nous nous sommes servis avec succès de cette nouvelle technique de télémétrie pour des scènes réelles afin de démontrer son potentiel pour la navigation d'un robot mobile.

Acknowledgements

I am pleased to thank all those who have helped me in large and small ways to complete this thesis. Leading this group are my co-supervisors, David Jones and Steve Zucker. David acted as my primary research advisor, and I have benefitted from a great deal of his time, enthusiasm, and ideas. Whatever quandary I found myself in, after a discussion with David, I usually found my way out. Steve provided direction at crucial points during my research, and his approach to computational vision has strongly influenced my thinking.

Thanks to all the people at the McGill Centre for Intelligent Machines who have made the last two years more enjoyable for me. In particular, fellow graduate students Pierre Breton, James Elder, and Mike Langer. Mike deserves special credit for reading an earlier version of this thesis and providing helpful feedback.

I am also indebted to Don Pavlasek for his mechanical expertise, Greg Dudek for the use of his mobile robot, and Kathy Murphy for her photographic equipment. I thank the Natural Sciences and Engineering Research Council of Canada for their financial assistance.

Finally, I would like thank all those who helped me to get here. Andrew Wong of the University of Waterloo first introduced me to computer vision. My friends and family provided much needed balance in my life. Most of all, I am grateful to my parents for their encouragement and support throughout my education.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The projection of a three-dimensional scene onto a two-dimensional photosensitive array is the foundation of visual perception in both man and machine. The goal of vision is, in part, to reconstruct the information apparently lost in this projection. In particular, *range imaging* is the process of computing the absolute depth of each point in a scene that is visible from a given viewpoint.

If projection occurs through an optical device with infinite depth of field, such as an ideal pinhole camera, quantitative depth information is completely lost. Cues such as perspective distortion, relative object size, and surface shading enable only the recovery of qualitative depth or surface orientation information.

However, if a scene is imaged from two slightly different viewpoints, either simultaneously or sequentially, depth may be reconstructed from binocular stereopsis or motion parallax, respectively. Both of these techniques require that information from two or more separate images be combined along a spatial or temporal axis. This requirement leads to the *correspondence problem*. A scene point appearing in one image need not appear similar (or even at all) in other images, therefore establishing a point-by-point correspondence between images is a difficult task. As the spatial or temporal separation between views is reduced, the correspondence problem becomes easier, but the depth information thus provided becomes less accurate.

In practice, the projection of a scene onto an image plane occurs through a de-

vice of *limited* depth of field, such as a finite aperture camera. The precise three-dimensional structure of the scene is mapped out between the lens and sensor plane as the locus of points where an image of the scene would be in exact focus. Furthermore, the depth of a point in the scene is related to the degree of *defocus* in its image, suggesting that a range image may be computed from a single intensity image. Unfortunately, without prior knowledge of the scene, changes in image intensity due to out-of-focus blur are not readily distinguished from those occurring in the scene itself. Therefore measurement of the amount of blur at a given image point is a difficult task, generally requiring multiple, identical views of the scene acquired with different depth of field.

The range sensing techniques described above are all *passive* techniques, in that they interpret visual images of a scene as it appears under ambient illumination. The advantage of passive techniques in computer vision is that they are general purpose, that is, applicable to a wide variety of scenes and viewing conditions. In many applications of range sensing, only limited control can be exerted over the scene, such as in aerial photography or stereomicroscopy. Perhaps what is most appealing about passive techniques is what they share in common with biological vision systems. Anyone doubting the capability of a passive vision system need only observe the ease with which we humans perceive the complex three-dimensional world around us, based solely on the images cast upon our retinae.

In contrast, *active* techniques rely on interacting with, in addition to observing, the scene. Instead of just the *scene acting upon the sensor*, in active vision the *sensor acts upon the scene*. These include non-visual techniques such as radar and sonar, as well as laser triangulation and other forms of structured lighting [9]. One such active technique is based on viewing a projected laser stripe with a conventional camera containing *two apertures* instead of the usual single aperture [10, 66]. The resulting camera image contains two laser stripes, one projected through each aperture. The local displacement between these two stripes, easily measured after some simple image processing, is related to depth in the scene. This range sensor is attractive in that it is monocular (requires only one image), inexpensive, compact in size and weight, and

relatively robust. The primary disadvantage is that depth information is provided only at the laser stripe positions in the image. To obtain a complete range image, the laser stripe must be swept across the scene, precluding a real-time sensor and requiring additional hardware. In this context, a real-time sensor is defined as a methodology that can be implemented in hardware to provide an output almost immediately upon receipt of the input.

This thesis describes a passive range imaging technique using a multiple-aperture camera, but one that can yield dense range images in real-time.

1.1 Motivation

Before the development of any range sensing system, it is important to consider the purpose for which the range image is to be used, and what criteria and constraints this imposes on the technique. Generally, there are two classes of applications for range images. The first requires high resolution, high accuracy range data for tasks such as object recognition and three-dimensional (3-D) model building. The second is more concerned with gross scene structure, such as the position and approximate shape of major objects in the scene, rather than fine surface detail. This type of data is often used for mobile robot navigation and obstacle avoidance. In this thesis the focus is on the latter of these two classes of applications.

Consider the sensory requirements of an autonomous mobile robot in an unknown, unstructured environment. In order for the robot to perform its task, it requires two kinds of information. First, it must have an approximate (though not necessarily complete) map of its environment that includes obstacles, walls, doorways, and other items of interest. Second, it must know its current position and orientation within this workspace. Given these two pieces of information, the robot can plan a path to its required destination and navigate along that path. However, due to the cumulative positional errors introduced by motorized locomotion, the robot should regularly confirm its position and orientation *while* in transit. This is particularly important in a cluttered or tightly spaced environment, where small positional errors can result

in catastrophic collisions.

The range sensing requirements in this type of application are very different from those in object recognition or 3-D model building. For example, suppose a chair is placed directly in the path of the mobile robot. It is irrelevant to the robot whether it is a four-legged or swivel type chair. What matters is that directly ahead there is a large "blob" of something, much closer than the background, around which the robot must manoeuvre. In fact, in mobile robotics there are practical concerns that dominate over the ability to make fine depth measurements. The size, weight, and power requirements of the range sensor may preclude a sophisticated laser range scanner mounted on a flexible arm. The need to transmit raw sensory data from robot to computer for processing leads to a preference for low resolution, monocular imaging devices over high resolution, binocular cameras. Most importantly, in order to integrate information from multiple views and update pose information while in motion, the processing time required to convert raw sensory data to 3-D range data is of paramount concern.

Taken together, the above arguments suggest a need in mobile robotics for compact, inexpensive range sensors that can make reliable, though not necessarily high resolution, 3-D measurements in real-time. This need motivates the approach taken in this thesis.

1.2 Overview

To appreciate the advantages of range sensing with a multiple aperture camera, the inherent difficulties with the conventional techniques of binocular stereopsis and depth from defocus are first described in Chapter 2. The use of a multiple aperture camera in an *active* vision system is also discussed. The problem of echo analysis, one that is central to this thesis, is introduced. Previous applications of echo analysis in computer vision are reviewed, and the problem of interpreting raw range estimates to generate an explicit representation of surfaces in a scene, is addressed.

The notion of *monocular stereopsis*, the computation of depth from a single image

formed with two apertures, is introduced in Chapter 3. The geometric optics underlying this principle are developed, in a manner that demonstrates the analogy between monocular stereopsis and depth from defocus. The formation of the *composite image*, the superposition of images from each aperture, is modelled as the sum of two identical but horizontally shifted images. This horizontal displacement varies with depth, and is analogous to the cue of horizontal disparity in binocular stereopsis. Therefore the term *monocular disparity* is introduced to refer to this displacement. Despite the apparent similarity with binocular stereopsis, it is shown that conventional solutions to the correspondence problem such as feature matching, phase-based methods, and correlation techniques either fail completely or have very limited success in measuring monocular disparity.

Since the formation of the composite image can be thought of as a *visual echo* process, a classical technique for echo detection, the *cepstrum*, is ideally suited to the monocular stereopsis problem. Unlike previous applications of the cepstrum to binocular stereopsis [81, 58, 48] and optic flow [4, 5], in monocular stereopsis the two images are already combined; the problem is to measure the echo between them, in effect, separating the two images. If the two images are distinct to begin with, there are many ways to determine correspondence (e.g., [22, 40]) which are not applicable to the monocular stereopsis problem.

In Chapter 4, the use of the cepstrum to estimate the monocular disparity over a composite image region of constant depth, is examined in detail. The classical notion of the cepstrum is refined to improve its robustness and efficiency, and the often ignored bias in the cepstrum resulting from echo truncation is analyzed. A major contribution of this thesis is the development of a model of the form of the composite image cepstrum, motivated by both mathematical and empirical results. This model explicitly describes how a visual echo manifests itself in the cepstrum, and how the underlying, unechoed image may obscure these echo cues. The model leads to a two-stage algorithm to measure monocular disparity from the cepstrum: a peak selection stage, and a peak localization stage. The effects of camera noise and out-of-focus blur on the performance of this algorithm are evaluated by quantitative experiments.

Finally, a confidence measure is derived that reflects the true distribution of errors in monocular disparity estimates. This distribution is a direct consequence of the two-stage algorithm for measuring monocular disparity from the composite image cepstrum.

The techniques developed in Chapter 4 are applied in Chapter 5 to transform a composite image into a representation of surfaces in the scene. The issues involved in computing a disparity map, such as the use of overlapping image windows and the selection of window dimensions, are addressed. The interpretation of this disparity map and the accompanying map of confidence measures, is considered as a visual surface reconstruction problem. For a given surface model, a maximum likelihood framework is developed to reconstruct surfaces in a scene based on monocular disparity estimates and the associated error distributions. A particular surface model, that of local planar facets, is used in this framework to generate an accurate representation of surfaces in a scene, even with many significant errors in the monocular disparity map.

It is important to realize that the system described in this thesis does not make measurements of depth at a single point in a scene, but over a region of the composite image. Furthermore, the output of the system consists of both the raw depth (or monocular disparity) measurements *and* an estimate of the error distribution for each measurement. This confidence measure is an integral part of the output, and cannot be ignored in evaluating the performance of this range sensor. Thus traditional techniques for evaluating the accuracy and resolution of range sensors [9] are not appropriate here. Instead, the framework of visual psychophysics provides a more general way to evaluate this and any other vision system, allowing a larger class of range sensors to be directly compared in quantitative terms.

The experimental techniques of human visual psychophysics are employed in Chapter 5 to evaluate the spatial resolution of this range sensor. Performance is evaluated in terms of an intelligent agent making two-alternative forced-choice decisions about its environment. The first experiment involves detection and discrimination of an obstacle of varying width. The second involves spatial localization of a step change

in depth. The results not only illustrate the better than expected spatial resolution of this range sensor, but also suggest how psychophysical methods may be applied to artificial as well as biological vision systems.

The method of passive monocular range imaging developed in this thesis is applied to a variety of real-world scenes in Chapter 6. These scenes are chosen to reflect different applications of range imaging, including the recovery of terrain structure, obstacle detection, locating objects for grasping, and robot navigation. Some concluding remarks are made in Chapter 7.

1.3 Contributions

The original contributions of this thesis are as follows:

- the application of cepstral analysis to the problem of computing depth from one composite image acquired by a multiple aperture camera
- a model of the form of the composite image cepstrum, consisting of the sum of: (1) the integer sampling of a waveform of triangle-shaped, alternating-sign peaks centered on integer multiples of the monocular disparity value, and (2) a discrete, uncorrelated, stationary, Gaussian noise process
- a technique for reliably identifying the peak of the cepstrum due to a visual echo occurring with a non-integer delay, given by the maximum pairwise sum of successive values of cepstrum
- given a composite image cepstrum, a maximum likelihood estimator of the precise monocular disparity, which exploits the entire cepstrum in a least-squares framework
- an estimate of the error distribution for a monocular disparity measurement provided by the cepstrum, given by a weighted combination of uniform and Gaussian distributions

- a technique to generate a piecewise planar representation of surfaces in a scene, based on estimates of monocular disparity and the associated error distributions, and a maximum likelihood criterion

- the application of techniques in human visual psychophysics to evaluate the spatial resolution of an artificial range imaging system

Chapter 2

Background

Rather than an exhaustive review of range imaging techniques in computer vision, this chapter will instead focus on several areas of research that are closely related to the work described in the body of this thesis. For a comprehensive review of active range sensors, the reader is directed to [9]. Passive range sensing techniques include structure from stereo, depth from defocus (both reviewed below), and structure from motion [80]. These should be distinguished from "shape from X" methods, where X is shading, texture, contour, etc., which provide surface orientation rather than absolute range data.

2.1 Binocular Stereopsis

Binocular stereopsis is perhaps the most popular method for passive range sensing in computer vision. When a scene is viewed from two slightly different locations, there are systematic differences or *disparities* between the two images that may be exploited to compute depth. The most salient of these image differences are the positional disparities of corresponding points. The difference in horizontal position between points in the left and right images that project from the same point in the scene (the horizontal positional disparity), allows the distance to the scene point to be inferred.

The difficult task in binocular stereopsis is to solve the correspondence problem, that is, to establish a point-by-point correspondence between the two images. Once the two images are brought into correspondence, positional disparities are easily measured, allowing the computation of a range image by triangulation. The correspondence problem is difficult because the two images are *not* simply shifted copies of each other. Since they are acquired independently from different viewpoints, there are different degrees of projective foreshortening, different photometric and optical properties, and different camera noise in the two images. Furthermore, if the scene contains abrupt changes in depth, there may be regions visible in one view that are occluded in the other.

Approaches to solving the correspondence problem attempt to overcome these difficulties by imposing on the matching process constraints that derive from physical properties of the scene and viewing geometry. The most commonly used are the *epipolar constraint*, which states that corresponding points lie on epipolar lines, and the *surface continuity constraint*, which assumes that disparities vary smoothly "almost everywhere" over the image [50]. The primary distinction among techniques for solving the correspondence problem lies in the type of primitive that is matched between images.

Feature-based schemes first extract a set of tokens from the two images, then match these tokens based on compatibility, uniqueness, continuity, and epipolar constraints. The features used include zero-crossings of oriented difference of Gaussian [51] or Laplacian of Gaussian [28, 62] filters, and linear edge segments [52, 2]. The algorithms used to obtain the best set of feature matches include relaxation labeling [35, 8, 45], dynamic programming [3, 57], and simulated annealing [7].

Another class of stereo algorithms attempts to match corresponding regions of the images themselves rather than features extracted from them. These are referred to as *area-based* techniques. They have the advantage of producing a dense depth map without the need for surface interpolation, but, as image intensities are less stable between views than edges, tend to be more susceptible to matching errors. Most area-based approaches use statistical measures such as normalized cross-correlation

or normalized sums of squared differences [25, 54, 23] to locate maximally similar image patches. A more sophisticated technique uses the differences in responses of a bank of orientation and spatial frequency tuned filters [38, 40]. These methods calculate, for each image patch in one image, a function that quantifies the similarity with image patches in the other image. The estimated positional disparity is given by the displacement between image patches, at which this similarity function attains its maximum value. There are several observations to be made regarding this technique which help motivate the approaches taken in Secs. 4.4 and 4.6.

The similarity function is continuous, but the images from which it is computed are discrete, so that the function is sampled at integer-valued (pixel) displacements only. It also tends to be slowly varying, since natural image intensities are locally correlated [70]. Therefore the site of the maximum of the similarity function is often approximated by the site of the maximum of the discretely sampled similarity function. Other techniques first identify the maximum of the discrete function, then use interpolation (based on a model of the peak shape) to more precisely estimate the location of the true maximum [23]. What is the distribution of error in disparity estimates using this technique? If the correct peak is selected, it may be assumed that disparity error (due to imperfect interpolation only) is Gaussian distributed. However, if the selected peak is in fact a *spurious* peak, not indicative of the true disparity, the estimated disparity may be radically different from the true disparity. Therefore it is incorrect to assume that all errors in binocular disparity measurements are Gaussian distributed.

The third class of approaches to binocular stereopsis which has of late received much attention is referred to as *phase-based* stereo [68, 37]. Although they may be referred to as area-based, these techniques measure positional disparity as a local phase difference between band-pass versions of the two images (local amplitude differences are discarded). This approach has the advantage that disparity can be measured directly to sub-pixel precision, without requiring the calculation of an explicit similarity function, or a peak selection and localization procedure. It has also been shown that phase information is more stable than amplitude under the deformations typ-

ical between left and right stereo images (e.g., changes in scale and contrast) [22]. Despite this observation, amplitude information may still be helpful in solving the correspondence problem. Indeed, systematic differences between left and right views, such as orientation and spatial frequency disparities, may be *exploited* in solving the correspondence problem, rather than simply treated as noise [38, 41].

This apparent paradox illustrates the dilemma of binocular stereopsis. It is the *differences* between the two views that *both* provide information about the 3–D structure of the scene, and make the correspondence problem a hard one. As the baseline or separation between the two viewpoints is decreased, the magnitude of these differences is reduced, but the triangulation upon which stereopsis is based becomes less accurate. In the extreme, the two views are identical making the correspondence problem trivial, but providing no depth information.

Some of the differences between two stereo views are unwanted, in that they convey no information about the 3–D structure of the scene, but make the correspondence problem more difficult. They include differences in focal length, zoom level, iris diameter, optical axis alignment, and lens distortion. These differences may be alleviated by, instead of taking one image from each of two cameras, taking two images from one camera. In this method, the scene must be static over the interval between taking the two images, and the camera must be moved to a second viewpoint or mirrors rotated within the camera [77], during this interval. Another solution is to take one image with one camera, but use an arrangement of mirrors such that this one image actually contains two stereo images, side by side [26]. This requires precise and somewhat awkward mirror and camera positioning. Nonetheless, provided the technical difficulties can be overcome, there are clearly advantages to a single camera stereo system.

2.2 Depth from Defocus

In practice there is no such thing as an ideal pinhole camera. Generally the level of illumination in a scene is such that a large iris diameter (compared to a pinhole)

is required to obtain sufficient brightness and contrast in the camera image. In this scenario, points at different depths are imaged with different degrees of focus. If it were possible to measure the precise amount of out-of-focus blur at each image point, geometric optics gives a simple expression for viewing distance [61] allowing direct determination of a range image from one intensity image. This principle is referred to as *depth from defocus*, as opposed to *depth from focusing*, which determines the sharpest of a sequence of images taken at different focal settings or viewing distances [46, 55].

Most approaches to depth from defocus model the out-of-focus image as the result of convolving the focused image with a blurring kernel, the size of which varies over the image. Geometric optics predicts the blurring kernel is a two-dimensional uniform function assuming the shape of the camera aperture (often called a pillbox function). For a circular aperture, this implies a point of light is blurred into a uniform intensity disc, whose diameter characterizes the amount of blur. However, due to the smoothing effects of diffraction, lens aberration, and the image digitization process, the actual blurring kernel often resembles a 2-D Gaussian function [61], characterized by its spread parameter σ_b. The uncertain relationship between σ_b and the ideal blur circle diameter (from which depth may be computed) is one of the shortcomings of this model of out-of-focus blur. For many cameras, neither the pillbox nor the Gaussian is a good approximation of the blurring kernel [19].

Assuming an appropriate model of the blurring kernel, the depth from defocus problem reduces to one of deconvolution. Given a blurred image patch (assumed to have constant depth throughout) representing the convolution of the focused image with some blurring kernel, the goal is to recover the unknown spread parameter of the blurring kernel. However, if the focused image is not available, the problem is underconstrained. It is impossible to distinguish changes in image intensity due to blur, from those due to the scene itself.

One solution to this problem is to analyze blur only in regions of the image where the scene properties are known, such as around intensity edges [31, 61, 74, 47]. However, this technique assumes that intensity edges in the focused image are perfect

step edges. Due to surface markings, spatially varying illumination, and camera noise, this is an unlikely scenario. In some applications, the characteristics of objects in the scene and the nature of the blur are fully known, so useful depth information is obtainable from a single image [36]. A more general solution is to acquire two identical images of the same scene, one with a pinhole camera to represent the focused image, another with a limited depth of field [61, 60]. Due to the illumination requirements of a pinhole camera, a more practical approach is to acquire two identical views with different, finite depth of fields [61, 76, 14], or focused at different depths [75]. The ratio in Fourier power between corresponding patches in the two images is then monotonically related to depth in the scene. A more recent method uses a matrix-based regularization approach that is independent of the functional form of the blurring kernel and less prone to windowing and border effects [19], although it is computationally expensive.

Although depth from defocus is often called a monocular range imaging technique, in practice two or more identical images are needed, acquired with different camera settings. These images may be acquired with multiple cameras, one camera with multiple, separate image planes, or one camera that takes multiple shots of the scene. Given these requirements, depth from defocus is, in practice, a binocular range imaging technique.

2.3 Depth from Multiple Apertures

Most cameras have one iris, a circular aperture that can be varied in size to vary the amount of light falling on the image plane. The larger the aperture, the smaller the depth of field (the range of viewing distances over which the image is in focus). For a point in the scene that is in focus, the cone of rays emerging from the scene point and passing through the aperture all converge at one point on the image plane. For a point that is out-of-focus, rays that pass through *different* parts of the aperture land on *different* parts of the image plane. The cone of rays passing through the aperture therefore forms a blur circle on the image plane. The diameter of this blur

circle encodes depth. If the iris is replaced by a mask with *two pinhole apertures*, this cone of rays is occluded at all but two points. The rays passing through the two pinhole apertures therefore form two separate points on the image plane. The distance between these two points encodes depth.

For example, suppose the scene is very simple, consisting of a point light source at some unknown depth. With a single aperture camera, as the depth at which the camera is focused varies from one extreme to the other, the image of the source will move into and then out of focus. When the camera is focused at a fixed depth, the image of the source is a disc, the diameter of which allows the depth of the source to be recovered (depth from defocus). With a double aperture camera, as the depth at which the camera is focused varies from one extreme to the other, the image of the source appears as two converging, coincident, and then diverging points of light. When the camera is focused at a fixed depth, the image of the source is two points of light, their *separation* allowing the depth of the source to be recovered (*depth from multiple apertures*).

This principle has been exploited to develop a compact active range sensor [66, 10, 64] known as "BIRIS" (meaning binocular iris). The sensor consists of two components — a double aperture camera and a laser stripe projector. The laser stripe is projected onto the surface of interest. The scene is viewed with a conventional CCD (charge coupled device) camera having a double aperture mask inserted in front of the lens or in the iris of the camera. The mask is aligned so that the two apertures lie on a line perpendicular to the orientation of the laser stripe in the image. For convenience, the stripes are projected parallel to the columns of the CCD array, and the apertures aligned parallel to the rows. An optical filter can be used with the camera to pass only wavelengths of light similar to the laser stripe. Therefore each scanline of the camera image consists of two peaks in intensity, corresponding to the two views of the laser stripe, which are identified by applying a one-dimensional smoothed derivative operator. Sub-pixel precision is obtained by interpolating the locations of the two zero-crossings. The resulting separation between laser stripes is converted to depth, yielding not a depth map, but a depth profile along one column of the image. To

obtain a dense depth map, the laser stripe must be actively swept across the scene so that a depth profile is obtained for each column.

Previous attempts to develop a *passive* multiple aperture range sensor had only limited success [66]. Using ambient lighting only, regions of the image acquired by a double aperture camera were analyzed by autocorrelation in an attempt to measure the separation between the views from each aperture. This technique is successful only for very highly textured scenes, such as provided by the projection of a laser speckle (random-dot pattern) into the scene. Range data provided by this sensor was reported to have "promising" resolution and accuracy, but the technique was abandoned in favour of the active technique described above.

An idea related to depth from multiple apertures is the plenoptic camera [1]. Here, instead of two pinhole apertures, a lenticular array is placed in front of the image plane to obtain depth information from a single shot. Each lenticule acts like a tiny pinhole camera, creating a macropixel representing an image of the scene as seen from some location within the image plane. From the set of all these macropixels it is possible to obtain different virtual viewpoints by selecting a particular pixel from each macropixel. The displacement between corresponding points in these views allows computation of a depth map, as in conventional binocular stereopsis. Unlike the multiple aperture camera described above, in the plenoptic camera the multiple views are not superimposed. Instead, the lenticular array simulates many cameras in one. In addition, this technique requires more specialized hardware than simply inserting a multiple aperture mask into a camera lens.

2.4 Echo Analysis and the Cepstrum

Many problems in early vision involve the analysis of repeating patterns in time or space. These include stereopsis, motion, texture, and symmetric boundary analysis. Such patterns may be considered as *echoes*, the superposition of repetitions of some underlying signal, separated by temporal or spatial delays. Echo detection and removal is a fundamental problem in signal processing and has applications in a wide

variety of fields. In general echo analysis, the delay of the echo does not exceed the length of the underlying signal, so that a portion of the signal and its echo overlap. Therein lies the challenge of echo detection. Since only the sum of the signal and its echo is observed, there is no obvious way to identify where the echo starts. The problem is similar to depth from defocus with a single image. Without prior knowledge of the unechoed signal, it is difficult to distinguish the original signal from its echo.

In signal processing the standard tool for analysis of echoes is the *cepstrum* [12]. The motivation and mathematics of the cepstrum are described in Sec. 4.1. For now, consider the cepstrum as a nonlinear system which takes as its input a composite signal consisting of the superposition of a signal and its echo, and outputs the delay between them. The cepstrum has been used extensively for echo detection in seismology [12], vocal pitch determination [56], decomposition of brain waves [44], and many other areas [17]. These applications have consistently shown the cepstrum to be effective on a broader class of signals and to be more immune to the effects of noise and distortion than other echo detection methods.

Cepstral techniques have also been applied to the binocular stereopsis problem [81, 58, 48] and visual motion analysis [4, 5, 6]. In these applications the signal and its echo are already separated (i.e., two or more distinct images are available). The goal is to measure the displacement between them. Therefore an initial step is required to combine windows from already separate images, to form a composite signal for cepstral analysis. The echo is appended to the end of, rather than superimposed on top of, the original signal. Nonetheless, these applications have shown the cepstrum to be an effective tool in the analysis of echoes in natural images.

2.5 Visual Surface Reconstruction

The data provided by many passive range sensing techniques is sparse, in that depth is provided only at scattered points throughout the visual field. The process of computing an explicit representation of surfaces in the scene that implicitly fills in these missing depth values is referred to as *surface reconstruction*.

Early attempts at surface reconstruction were based on minimizing quadratic variation in surface orientation between the locations of zero-crossings (where depth was provided by binocular stereopsis) [29]. Since there is often uncertainty in the depth measurements themselves, a second term was added to the objective function, given by the weighted squared error between given depth measurements and the reconstructed surface. Minimization of this functional has become known as the *thin plate spline* technique. The result is a unique, C^1 continuous surface. This technique has several major drawbacks. First, the assumption that the scene consists of a single C^1 surface is often a poor one. Because the solution surface is smooth, it tends to oscillate on either side of depth discontinuities in the scene, while at the same time blurring the actual discontinuities themselves. The degree of smoothness in the solution is controlled by an arbitrarily chosen constant, the weight of the "fit-to-data" term relative to the "smoothness" term in the objective function. Finally, in practice the minimization procedure is slow to converge. A sophisticated multi-level relaxation technique may lead to faster convergence [78].

Various adaptive schemes for discontinuity preservation have since been proposed, such as statistical hypothesis testing on the parameters of locally fitted planar patches [30], and detection of high surface bending in the vicinity of inflection points [79]. Another technique suggests that the surface be allowed to crease or fracture whenever the energy so released is worth paying an extra penalty [11]. This leads to the minimization of a function containing multiple local minima, which is solved using a *graduated non-convexity* algorithm. A similar non-convex minimization problem arises in computing a *maximum a posteriori* (MAP) estimate of an original image given the degraded image [24]. Based on knowledge of the degradation processes and a Markov random field image model, the MAP image estimate is computed using a *simulated annealing* technique. In practice, simulated annealing is also slow to converge.

If the scene is well approximated by one smooth surface, all of these surface reconstruction algorithms perform well *if* the range data is corrupted only by uncorrelated Gaussian noise. If some data points have higher confidence than others, different

estimates of standard deviation of noise can be associated with different points. This leads to weighted least-squares approaches where the more uncertain a measurement, the lower its contribution to the fit-to-data term. However, errors in range data provided by passive techniques are rarely Gaussian, particularly in the case of binocular stereopsis. If a matching error is made, not only may the resulting disparity be dramatically different from the true value, but it is likely that this error occurs over a neighbourhood rather than at a single isolated point (since matching errors are often due to image structure over a region). Therefore a clump of incorrect disparities may be interpreted as valid surface structure. A rapid change in depth may be interpreted as a valid surface discontinuity, which some surface reconstruction algorithms will obediently try to preserve. What is needed is first a realistic model of the distribution of errors in the range data, and then a surface reconstruction technique that exploits the estimated parameters of this model at each depth measurement. This is provided in Secs. 4.6 and 5.2 respectively.

Chapter 3

Monocular Stereopsis

Binocular stereopsis refers to the ability to compute depth from the differences in two views of a scene taken from different viewpoints. It is inherent to this paradigm that two separate images are available for analysis. As described in Sec. 2.3, it is possible to sense depth from one image consisting of two superimposed views acquired through separate apertures. A new term is introduced to refer to this principle — *monocular stereopsis*, literally meaning "solid sight with one eye". Like binocular stereopsis, depth is recovered from the correspondence of two views, however, in monocular stereopsis this correspondence is determined *within one composite image*, instead of *between two separate images*.

In this chapter the equation allowing the computation of depth in monocular stereopsis is developed. A model is presented to describe the formation of the composite image from the image seen through one pinhole aperture. This model forms the basis for the technique developed in Chapter 4 to solve the monocular stereopsis problem. The fundamental differences between solving the binocular and monocular correspondence problems will also emerge.

3.1 Geometric Optics

Consider a normal lens camera in which a mask containing two apertures is inserted in place of the iris. The apertures are identical in size and shape, and are equally spaced about the optical axis of the camera. Geometrically, the mask may be represented as being in the centre of the camera lens (see Fig. 3.1). The thin lens approximation is assumed to be an adequate model of the camera optics [34]. The relevant camera parameters are the focal length of the lens, F, the distance between the two apertures, D, the diameter of each aperture, A, and the distance from the lens to the sensor plane, f.

The well known Gaussian lens equation [34] gives

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{Z} \tag{3.1}$$

where Z is the distance from the lens to a reference plane in the scene, the image of which is in focus on the sensor plane. Consider a point $P(x_P, y_P, z_P)$ in the scene, forming an out-of-focus image on the sensor plane. Let f_P be the distance from the lens to the plane upon which the image P' of P is in focus. Notice that the images of P from each aperture are not only both in focus at P', but also coincide (see Fig. 3.1). The lens equation now gives

$$\frac{1}{F} = \frac{1}{f_P} + \frac{1}{z_P} \tag{3.2}$$

On the sensor plane there are two cues to the depth of P: the distance d_P between the images of P arising from each aperture, and the diameter a_P of each blur circle. The triangles with bases A and a_P and altitudes f_P and $f_P - f$ are similar, as are the triangles with bases D and d_P and altitudes f_P and $f_P - f$, giving

$$\frac{f_P - f}{f_P} = \frac{a_P}{A} = \frac{d_P}{D} \tag{3.3}$$

Substituting for f_P in Eqn. (3.2) from Eqn. (3.3) gives the depth z_P of P, in terms

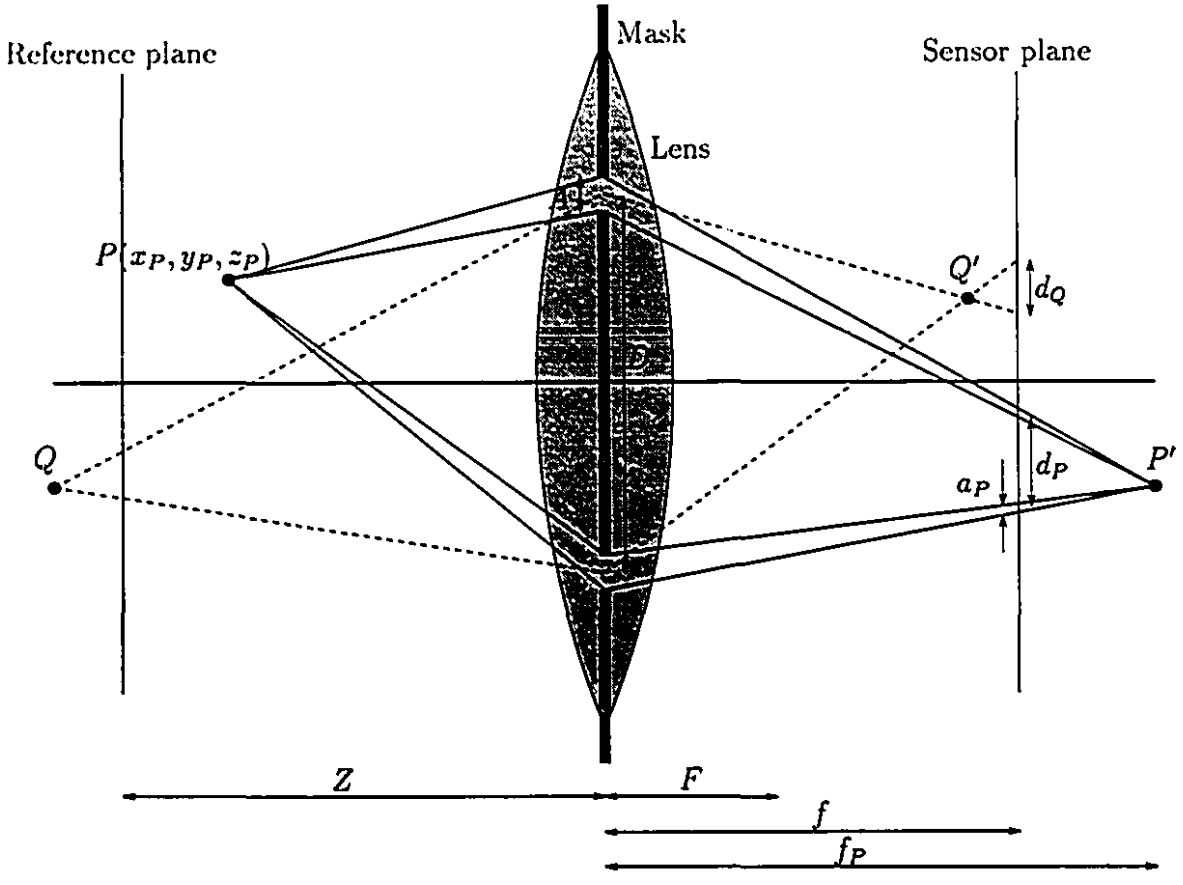


Figure 3.1: Geometric optics for a double aperture camera. A sensor plane is at a distance f from a lens with focal length F. A reference plane, conjugate to the sensor plane, lies at a distance Z in front of the lens, so that all points on the reference plane are imaged in focus. A mask containing two small apertures of diameter A separated by a distance D, is placed in the fully open iris of the camera. A point P, in front of the reference plane, has an image P' at a distance f_P from the lens, but two blurred images on the sensor plane separated by a distance d_P and with blur circle diameter a_P. Another point Q, beyond the reference plane, also produces two images, separated by a distance d_Q.

of the ratio of blur circle diameter to aperture diameter,

$$\frac{1}{z_P} = \frac{1}{F} - \frac{1}{f}\left(1 - \frac{a_P}{A}\right) \tag{3.4a}$$

or in terms of the ratio of image point separation to aperture separation,

$$\frac{1}{z_P} = \frac{1}{F} - \frac{1}{f}\left(1 - \frac{d_P}{D}\right) \tag{3.4b}$$

The first equation (3.4a) is the *depth from defocus* equation [61], where the range of an imaged point is calculated from its blur circle diameter relative to the camera aperture diameter. The second equation (3.4b) is identical except aperture diameter is replaced by distance between two apertures, and blur circle diameter replaced by distance between two images of the same point in the scene. This equation is the basis for monocular stereopsis. The distance d_P, the displacement between the two images of P, is referred to as the *monocular disparity* value.

It is important to appreciate the relationship between monocular disparity and depth. Monocular disparity is what can be measured from a composite image; depth is the desired end product of monocular stereopsis. A plot of depth versus monocular disparity for a particular camera configuration helps to provide some intuition for this relationship (see Fig. 3.2). In this example, the dashed-line curve is for the camera focused at a depth of 0.3 m; for the solid-line curve, the camera is focused at infinite depth. In terms of Eqn. (3.4b), the only difference between these two curves is in the value of f, which is responsible for the apparent shift between the two curves. Negative disparities correspond to depths greater than the depth at which the camera is focused (such as point Q in Fig. 3.1). Notice that at different points along the curves, the same change in disparity corresponds to very different changes in depth. For example, on the dashed-line curve, the difference in depth between disparities -23 and -24 pixels is 1.3 m, while the difference between disparities $+23$ and $+24$ pixels is 0.003 m. This nonlinear relationship has important implications for the interpretation of errors in monocular disparity estimates, and comes up frequently in Chapters 4, 5, and 6.

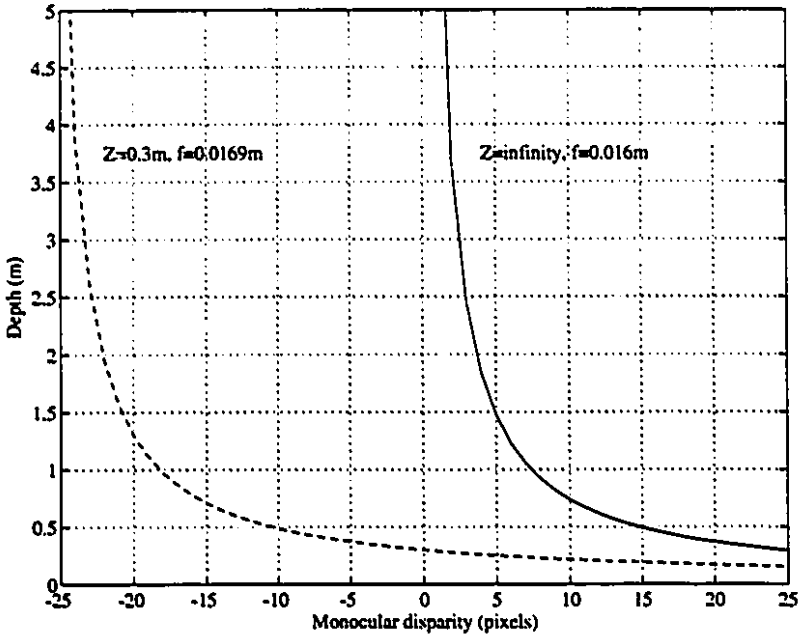
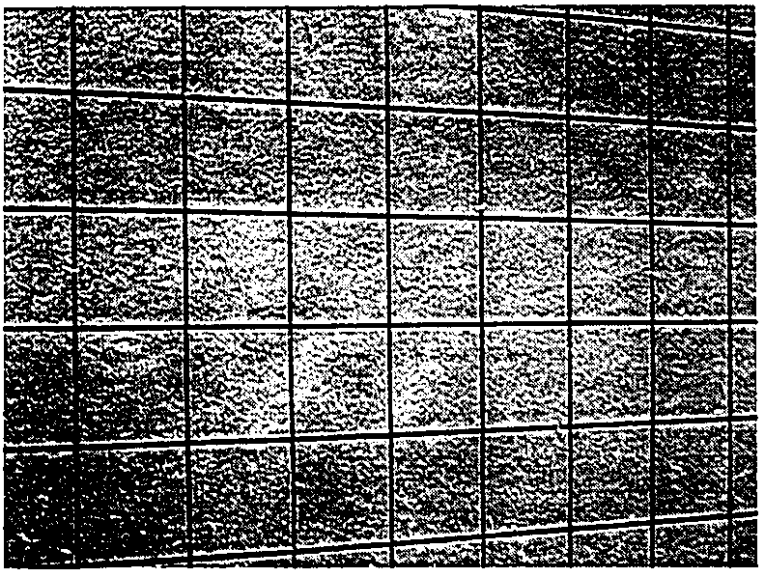


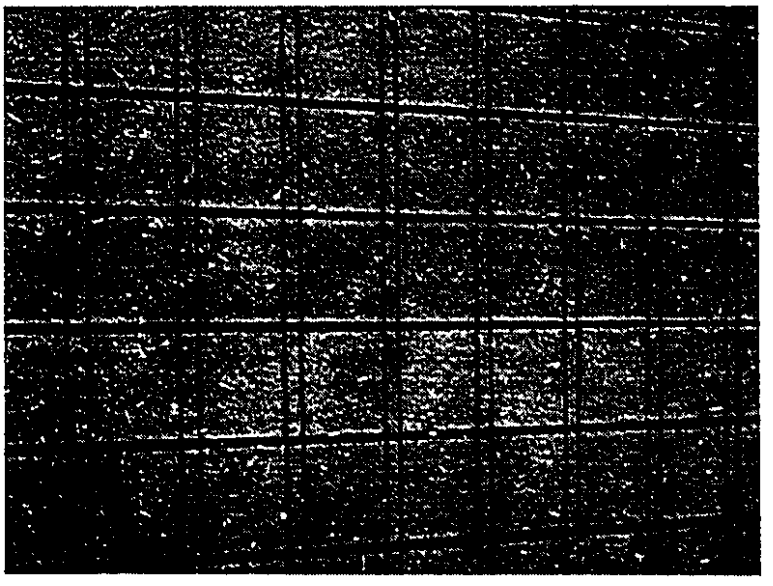
Figure 3.2: Depth versus monocular disparity. The depth of point in a scene, z_P, is plotted as a function of its monocular disparity, d_P, according to Eqn. (3.4b). In both cases, $F = 16.0$ mm, $D = 6.0$ mm, and the composite image is assumed to have resolution 640×480 pixels. For the dashed-line curve, the camera is focused at $Z = 0.3$ m, implying from Eqn. (3.1) that $f = 16.9$ mm. For the solid-line curve, the camera is focused at infinity, so that $f = F$.

In Fig. 3.1, the two images of P on the sensor plane are not points but rather discs, since each aperture as a finite diameter A. This blur can reduce the accuracy of monocular stereopsis, for it creates uncertainty in the measurement of d_P. If the two apertures were ideal pinholes, there would be no blur in the images of P and this problem would not occur. In practice, pinhole apertures are impractical due to the high scene illumination or exposure times required to obtain a composite image with sufficient brightness and contrast. The two goals of minimizing composite image blur and maximizing light admittance would appear to be contradictory. As a solution to this problem, non-circular apertures can be used. It is convenient in practice to rotate the double aperture mask so that the two apertures are aligned with the scanlines of the composite image. In this case, monocular disparity in the composite image has a horizontal component only. Therefore blur in the horizontal direction introduces much more uncertainty in monocular disparity estimates than blur in the vertical direction. Since the shape of the blurring kernel is roughly the shape of the apertures, the horizontal size of the two apertures should be minimized, while the vertical size is less critical. To admit the most light while minimizing blur in the horizontal direction, vertical slit shaped apertures can be used. With such apertures, vertical scene features appear sharp with an easily noticeable monocular disparity, while horizontal features are noticeably blurred (see Fig. 3.3).

One drawback to both monocular stereopsis and depth from defocus is the inherent ambiguity in the sign of the d_P and a_P. For example, in Fig. 3.1 there is no way of determining from the sensed image that the point Q is behind the reference plane rather than in front. As illustrated in Fig. 3.2, points behind the reference plane give rise to negative (crossed) disparities; points in front have positive (uncrossed) disparities. With the technique for estimating monocular disparity developed in Chapter 4, not only is the sign of disparity not recoverable, but very small disparities are difficult to detect, zero disparity being impossible. To resolve these potential difficulties, the images from each aperture may be diverged slightly, so even a point on the reference plane gives rise to a non-zero monocular disparity. This may be accomplished either by inserting a prism into the lens system [66], or by separating halves of a spherical



(a)



(b)

Figure 3.3: Single and composite image of a slanted plane. **(a)** An image of a plane slanted from left to right, taken with a single aperture camera. **(b)** A composite image of the same scene taken with a camera with two vertical slit apertures. Note that vertical lines remain relatively sharp, while horizontal lines appear quite blurred.

lens [65]. Points in front of or behind the reference plane then lead to positive or negative differences from a reference disparity value. If such specialized apparatus is not available, there is an inexpensive alternative. The camera may be focused at a point closer or farther than the entire scene to be observed, so that all disparities are the same sign and none are close to zero. In particular, the camera may be focused at infinity, as in the solid-line curve of Fig. 3.2, so that all disparities in the scene are positive.

3.2 The Composite Image

In order to solve the monocular stereopsis problem, a model of the composite image acquired by a double aperture camera is developed in this section. Taking a single aperture image patch as the input and composite image patch as the output, the double aperture imaging process is considered as a linear system with some unknown parameter d, the monocular disparity value. An enlarged portion of a composite image of a distinctive textured pattern (see Fig. 3.4) should provide the reader with some intuition for this model.

3.2.1 Spatial Domain Model

Since the two apertures in the iris mask are closely spaced and are identical in size and shape, the images acquired via each aperture are very similar. Over a region of constant depth, these images are assumed to be identical. The composite image formed on the sensor array is the sum of these two images. The apertures are equally displaced from the optical axis along a line parallel to the scanlines of the CCD array, so the displacement between the two views forming the composite image has a horizontal component only. Therefore over a window of constant depth, z, the composite image may be modelled as

$$c(x,y) = s(x,y) + s(x-d,y) \tag{3.5}$$

Figure 3.4: Composite image of a fronto-parallel plane. An enlarged portion of a composite image of a Canadian five dollar bill, placed flat on a plane fronto-parallel to the camera. The monocular disparity is approximately 13 pixels throughout the image.

where $s(x, y)$ is the single aperture image, and d is the monocular disparity value, related to z through Eqn. (3.4b). Therefore the composite image may be considered as the superposition of the single image and a shifted version of itself. In acoustics, the repetition of a signal after a temporal delay is referred to as an *echo*. In computer vision, the repetition of an image after a spatial delay is referred to as a *visual echo*. In monocular stereopsis, the spatial delay of the visual echo is the monocular disparity value.

Since the visual echo is an entirely horizontal phenomenon, the monocular stereopsis problem may be solved in one-dimension (1–D), that is, by computing monocular disparity independently within each scanline. In the case of vertical slit apertures, image data is significantly blurred in the vertical direction. Since neighbouring scanlines in the composite image are often very similar, this vertical blur tends to "compensate" for any small misalignment of the apertures with the image scanlines. Thus the monocular stereopsis problem may be solved one scanline at a time, or even better, all scanlines in parallel.

In 1–D, the formation of the composite image may be written as the convolution

of the single aperture image with two impulses separated by a distance d, that is,

$$c(x) = h_d(x) * s(x) \tag{3.6a}$$

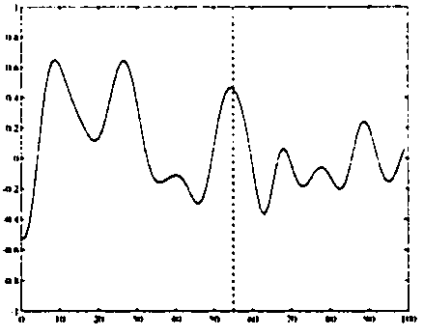
where

$$h_d(x) = \delta(x) + \delta(x - d) \tag{3.6b}$$

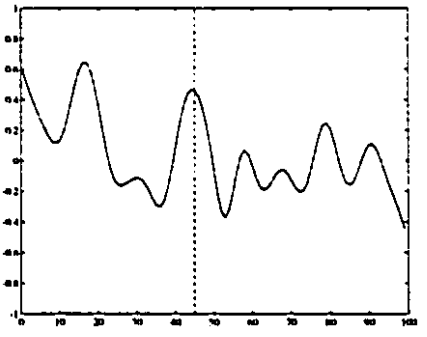
The problem then is to recover the system parameter d from the measured composite image $c(x)$. It is important to realize that there is no prior knowledge of the single aperture image $s(x)$. Clearly a simple solution would be to measure $s(x)$ with a single aperture camera, then $c(x)$ with a double aperture camera, and perform a system identification procedure to determine $h_d(x)$. This is analogous to the depth from defocus technique of obtaining two identical images of a scene, one with a pinhole camera, the other with a limited depth of field [60]. However, this solution requires two images, defeating the purpose of monocular stereopsis. Therefore the single image is assumed to be unavailable in solving the monocular stereopsis problem. This is a more challenging problem, since for a given $c(x)$ and *any* value of d, there exists an image $s(x)$ satisfying Eqn. (3.6), which can be recovered by deconvolution.

To further appreciate the implications of this model, an example of the formation of a composite signal from a single signal and its echo is presented. In this example, a discrete signal consisting of smoothed white noise (see Fig. 3.5a), is echoed with a delay of 10 sample points (see Fig. 3.5b). The dashed vertical line in these two plots indicates corresponding points, the signal structure around which is identical. The point-by-point sum of the original signal and its echo yields the composite signal (see Fig. 3.5c). In the composite signal, the structure around "corresponding" points (indicated by two dashed vertical lines) is no longer similar. To appreciate the difficulty of echo analysis, cover the top two curves and try to estimate the echo delay from the composite signal alone.

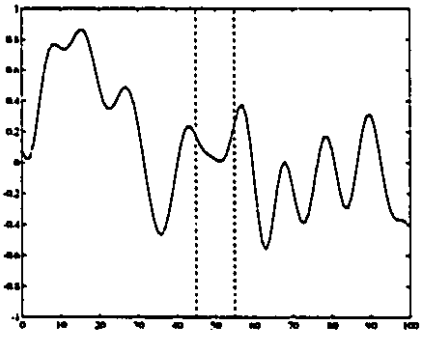
In previous work [39], we sought to exploit the characteristics of natural images, such as bounded contrast and spatial frequency, to evaluate the *feasibility* of a reconstructed single image for different candidate values of d. For a discrete, finite length



(a)



(b)



(c)

Figure 3.5: Formation of a composite signal. **(a)** A smoothed, discrete random signal. **(b)** The echo of the signal in (a), with a delay of 10 sample points. The dashed line indicates the position of two corresponding points. **(c)** The composite signal, given by the pointwise sum of (a) and (b). Note that corresponding points no longer exhibit similar structure.

$c(x)$, the estimated single aperture image, $\hat{s}_d(x)$, may be recovered by the matrix multiplication

$$\hat{s}_d = H_d^* \, c \tag{3.7}$$

where H_d^* is the precomputed pseudo-inverse of the matrix representation of $h_d(x)$, for some candidate disparity value d. The feasibility of $\hat{s}_d(x)$ is measured by its normalized contrast relative to the given composite image. Although this technique was successful in many experiments and may be implemented as an efficient 1-D recursive inverse filter [39], there is no guarantee that incorrect disparity estimates will *not* lead to maximally feasible $\hat{s}_d(x)$ signals, and therefore the technique is not robust.

3.2.2 Frequency Domain Model

In the frequency domain, the composite image is modelled as the product of the Fourier transform of the single aperture image, $S(\omega)$, and the echo process transfer function $H_d(\omega)$, that is,

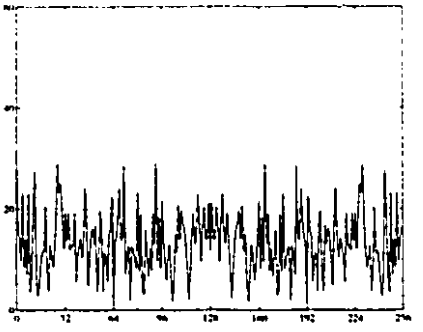
$$C(\omega) = H_d(\omega)S(\omega) \tag{3.8a}$$

where

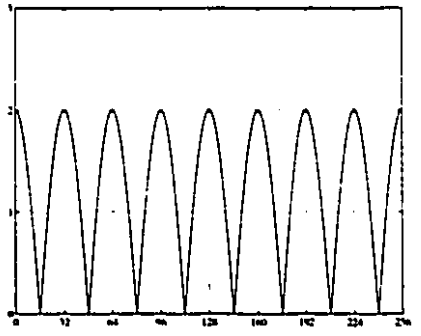
$$H_d(\omega) = 1 + e^{-j\omega d} \tag{3.8b}$$

Similar to the spatial domain deconvolution in the previous section, for a given $C(\omega)$ and *any* value d, the Fourier transform of the single aperture image, $\hat{S}(\omega)$, may be recovered directly from Eqn. (3.8). Since natural images are known to contain significantly more power at low frequencies than high frequencies [21], a feasibility measure may be developed to select the most likely reconstructed single image spectrum out of a range of candidates [39]. However, this technique will suffer from the same lack of robustness as the corresponding spatial domain technique described above.

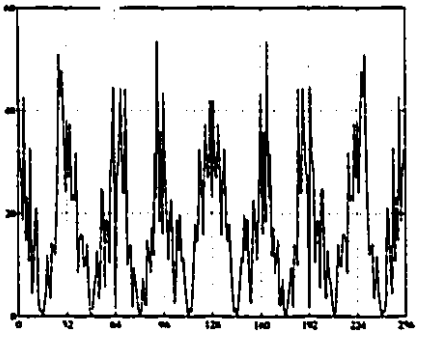
Before moving on it is instructive to examine qualitatively the effect of the visual echo process on the Fourier spectrum of the composite image. When a single image, $s(x)$, is convolved with the impulse response $h_d(x)$, its Fourier transform, $S(\omega)$, is



(a)



(b)



(c)

Figure 3.6: Formation of a composite signal spectrum. **(a)** The spectrum of a zero-mean, unit variance, Gaussian white noise signal. **(b)** The magnitude of the echo transfer function, $H_d(\omega)$, for $d = 8$. There are d ripples over the discrete spectrum, hence the "frequency" of this cosinusoid (in the frequency domain) is d. **(c)** The resulting composite signal spectrum, given by the product of (a) and (b). Note the composite signal spectrum exhibits the same frequency ripple as the echo transfer function, but it is partially obscured by the spectrum of the underlying signal.

multiplied by the cosinusoidal transfer function $H_d(\omega)$. This leads to attenuation of certain frequencies in the composite image spectrum, or a *ripple* in $|C(\omega)|$ (see Fig. 3.6). For an echo of delay d, a ripple of "frequency" d appears in the spectrum of the composite image, however, this ripple is partially obscured by the spectrum of the underlying single image, $|S(\omega)|$.

3.2.3 Incorporating Blur and Noise

Although Eqn. (3.5) expresses the relationship between the composite image and an image from one aperture, it is not a complete model of the image acquired by the double-aperture camera, for it ignores the effects of out-of-focus blur and camera noise. The extent to which blur and noise affect estimates of monocular disparity may place constraints on the quality of optics and image acquisition hardware required for monocular stereopsis. These effects are examined in Sec. 4.5.

If $s(x,y)$ is the noise-free single image as seen through one ideal pinhole aperture, d the monocular disparity value which varies with depth in the scene, a the diameter of the blur circle which also varies with depth, and $n(x,y)$ a noise field, the composite image over a region of constant depth can be expressed as

$$c(x,y) = B_a(x,y) * s(x,y) + B_a(x,y) * s(x-d,y) + n(x,y) \tag{3.9}$$

where $B_a(x,y)$ is the blurring kernel, assumed to be identical for both apertures. Ideally the blurring kernel assumes the shape of the aperture, so for a circular aperture the operator is a circular "pillbox" of diameter a,

$$B_a(x,y) = \begin{cases} \frac{4}{\pi a^2} & \text{for } x^2+y^2 \leq \frac{a^2}{4} \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

Due to the combined effects of diffraction, lens aberration, and image digitization, the blurring kernel can be more realistically modelled as a 2-D Gaussian function [61, 74, 76],

$$B_a(x,y) = \frac{1}{2\pi r a^2} e^{-\frac{1}{2}\frac{x^2+y^2}{ra^2}} \tag{3.11}$$

where r is some camera dependent constant (often approximated by 1/2 [74]). An alternative model which generalizes to other apertures is to consider the blurring kernel as the convolution of an aperture shaped pillbox with a 2–D Gaussian. Such a model can well approximate blurring kernel of a camera with vertical slit apertures.

The camera noise added to the composite image is modelled as uncorrelated, zero-mean, Gaussian distributed, with standard deviation σ_n. In Sec. 4.5, monocular disparity is measured under increasing levels of artificially generated noise, to determine the signal-to-noise (SNR) rating required of a camera in order for monocular stereopsis to be successful.

Exploiting linearity of the convolution operator, the model of composite image formation in Eqn. (3.9) may be written as

$$c(x,y) = s(x,y) * h_d(x,y) * B_a(x,y) \; + \; n(x,y) \tag{3.12}$$

where $h_d(x,y)$ is the echo impulse response given by Eqn. (3.6b).

3.3 Inappropriateness of Conventional Stereo Methods

The problem in monocular stereopsis is to recover the monocular disparity, or displacement between the two single aperture images, at each pixel in the composite image. This may seem very similar to the binocular correspondence problem, where disparity is determined between corresponding points in two separate images. However, because the two images are superimposed and only the composite image is available, the monocular correspondence problem is very different. The information that is trivial in the binocular case — knowing which image data is due to which of the two views (often called *eye of origin* information in biological vision) — is completely lost. To understand the implications of this loss, the manner in which conventional binocular stereopsis methods break down when applied to monocular stereopsis is examined in this section.

3.3.1 Feature-based Techniques

Many binocular stereo algorithms are based on identifying what are thought to be stable features (such as edges) in each image and then matching compatible features between images along epipolar lines [51, 28, 3, 62]. To apply such a technique to the monocular stereopsis problem, all features would be identified in the composite image, and those aligned with the two apertures matched according to some ordering constraint. But what comprises a stable feature in a composite image? The appearance of each composite image feature is always given by the sum of two images, and for "corresponding" features, one component of this sum in each occurrence will be different. In other words, matching composite image features is like trying to match $a + b$ and $b + c$, where $a \neq c$ and the relative magnitudes of a, b, c are unknown. Two corresponding features in the composite image may be *arbitrarily* different (see Fig. 3.5). Therefore, feature-based stereo matching schemes are inappropriate for the monocular stereopsis problem.

3.3.2 Phase-based Techniques

Another class of stereo algorithms is based on measuring local phase differences between the outputs of band-pass filters applied to the left and right images [37, 22]. Such a technique is not applicable to the monocular stereopsis problem. When two identical sine waves with some constant phase difference are added together, the result is a new sine wave, whose phase reveals nothing about the original phase difference.

One could assume the band-pass version of the composite image is given by the sum of two band-pass signals with some constant local phase difference. In other words, the output of a Sine Gabor filter applied to the composite image may be modelled as

$$c_{sin}(x) = \rho \sin(\bar{\omega} x + \phi_1) + \rho \sin(\bar{\omega} x + \phi_2) \tag{3.13}$$

where $\bar{\omega}$ is the peak pass frequency of the Gabor filter, ρ is the amplitude of the Gabor response of the single image, and $\phi_2 - \phi_1$ is the phase difference from which the monocular disparity value may be calculated. Using trigonometric identities this

expression may be rewritten as

$$\begin{aligned} c_{sin}(x) &= 2\rho \cos\left(\frac{\phi_2 - \phi_1}{2}\right) \sin\left(\bar{\omega}\, x + \frac{\phi_1 + \phi_2}{2}\right) \\ &= \rho' \sin(\bar{\omega}\, x + \phi') \end{aligned} \tag{3.14}$$

There are two measurable quantities from the composite image, ρ' and ϕ'. Without prior knowledge of the single image, there are three unknowns: ρ, ϕ_1 and ϕ_2. Therefore the problem is underconstrained, and the phase difference $\phi_2 - \phi_1$ is not recoverable from the composite image.

Eqn. (3.13) provides some insight into the nature of the composite image Fourier spectrum. For a fixed monocular disparity, different frequencies $\bar{\omega}$ lead to different phase differences $\phi_2 - \phi_1$. At some frequencies, the phase difference is such that peaks and troughs of the two sinusoids are aligned, so that the amplitude of the resultant sinusoid is minimized, while at other frequencies, peaks and peaks are aligned, so the resultant amplitude is maximized. Therefore in the composite image some frequency components are amplified while others are attenuated, leading to a ripple in the composite image spectrum (see Fig. 3.6).

3.3.3 Correlation Techniques

A third class of stereo algorithms uses area correlation techniques to locate maximally similar image patches between views [25, 54, 23]. When applied to monocular stereopsis, cross-correlation between two images becomes autocorrelation within one image. Initially, autocorrelation seems like an appropriate technique to estimate monocular disparity. One would expect the inner product of composite image patches separated by the monocular disparity to be significantly larger than that for other lags. However, composite image patches separated by the monocular disparity need not be similar; in fact they may be arbitrarily different (see Fig. 3.5).

A similar argument in which autocorrelation appears to be a solution to monocular stereopsis, but in fact is not, can be made in the frequency domain. For a signal that has no imaginary component, autocorrelation may be defined as the Fourier

transform of the power spectrum of the signal. The power spectrum of the composite image contains a ripple with "frequency" equal to the monocular disparity, so *its* Fourier transform should contain more power at the monocular disparity value than at other "frequencies". Therefore the autocorrelation function of the composite image is expected to contain a peak at the correct disparity value, and thus serve as a solution to the monocular stereopsis problem. However, the power spectrum of the composite image contains ripples due to both the visual echo process *and* the single image, as depicted in Fig. 3.6. It is not clear in this case that the echo ripple will dominate over these other ripples.

To further investigate the performance of autocorrelation in echo detection, experiments were performed with artificially generated signals. Gaussian distributed white noise was used as the single signal, and echoed by a known delay. The normalized autocorrelation function of the composite signal was then computed. To simulate natural imagery, which are known to contain more energy at low frequencies than high frequencies [21], the single signal was low-pass filtered with a decreasing cutoff frequency. The experiment was repeated with a series of randomly generated inputs. A typical result is presented in Fig. 3.7.

When the single signal is not low-pass filtered (i.e., it is white noise), there is usually a strong peak in autocorrelation at the correct echo delay (see Fig. 3.7a,b). Since the spectrum of a white noise signal is uniform across all frequencies, the ripple due to the echo is quite apparent. However, as the single signal is low-pass filtered, this peak in autocorrelation decreases in height (Fig. 3.7c-f), eventually becoming submerged in noise (Fig. 3.7h).

These results suggest that unless the single image of the scene is white noise, autocorrelation is an unreliable means of estimating monocular disparity. Images of real-world scenes under ambient illumination very seldom resemble white noise [21]. One solution to this problem is to use an active form of illumination, such as a laser speckle projector, to ensure that surfaces in the scene do appear as white noise [66]. Another solution is to use a method of visual echo analysis that is less sensitive to ripples in the single image spectrum. This is described in Chapter 4.

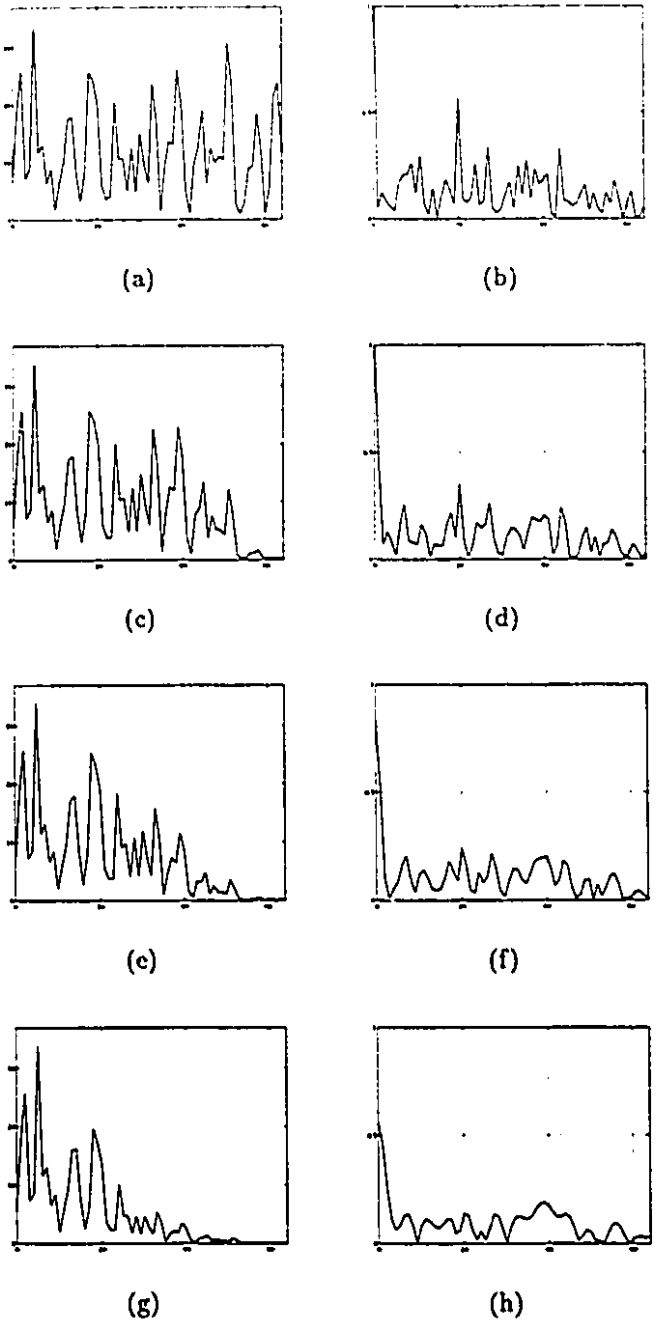


Figure 3.7: Autocorrelation as a means of detecting monocular disparity. **(a)** The Fourier spectrum of a composite signal, generated from an original signal consisting of white noise, and echoed with a delay of 20 sample points. **(b)** The autocorrelation function of the composite signal whose spectrum is given in (a). The echo delay of 20 is successfully detected. **(c,e,g)** The same Fourier spectrum as in (a), low-pass filtered with a decreasing cut-off frequency. **(d,f,h)** Autocorrelation functions of the signals whose spectra are given in (c,e,g) respectively. The peak at the correct echo delay of 20 decreases in height, until it becomes indistinguishable from noise.

Chapter 4

Cepstral Analysis of the Visual Echo

The problem of monocular stereopsis may be formulated as measuring the delay of the visual echo at each point in a composite image. This chapter develops a reliable technique to estimate this spatial delay (or monocular disparity) over a region of the composite image with constant depth in the scene. The technique is based on the *cepstrum*, a tool used in signal processing to detect and analyze echoes. The cepstrum is more reliable than autocorrelation for estimating the delay of an echo because it is less sensitive to the structure of the single image. A model of the composite image cepstrum is proposed, which leads to an algorithm for estimating monocular disparity to sub-pixel precision, and a confidence measure for each such estimate. These estimates and confidence values are used in Chapter 5 to compute a higher level representation of surfaces in the scene.

4.1 The Cepstrum

The spectrum of a composite image contains a ripple due to multiplication of the single image spectrum by the echo transfer function ripple. The frequency of this ripple (in the frequency domain) is precisely the echo delay d, the monocular disparity value to be recovered. However, for signals that are non-white, the Fourier transform of the power spectrum (the normalized autocorrelation function) is not a reliable detector

of this ripple frequency. Adopting the symbols used in Sec. 3.2.2, the power spectrum of the composite image is given by the squared magnitude of Eqn. (3.8), that is,

$$|C(\omega)|^2 = |H_d(\omega)|^2 \, |S(\omega)|^2 \tag{4.1}$$

Therefore, from the standpoint of identifying the ripple frequency d, $S(\omega)$ acts as multiplicative noise. It is common in signal processing applications to use a nonlinear operator to transform multiplicative noise into additive noise, so that linear filtering may be used to separate signal from noise [59]. Taking the logarithm of Eqn. (4.1) gives

$$\log |C(\omega)|^2 = \log |H_d(\omega)|^2 \; + \; \log |S(\omega)|^2 \tag{4.2}$$

which transforms $S(\omega)$ into additive noise, so that a subsequent linear operator (the Fourier transform) is better able to identify the ripple of $H_d(\omega)$. The Fourier transform of Eqn. (4.1) is the *convolution* of "signal" (from the visual echo) and "noise" (from the single image); the Fourier transform of Eqn. (4.2) is the *sum* of "signal" and "noise". In general, noise has a more detrimental effect when convolved with a signal, than when added to the signal.

The procedure of computing the power spectrum of a given signal, taking its logarithm, and computing the power spectrum of the result, is referred to as taking the *power cepstrum* of the signal [12]. In other words, the power cepstrum is the power spectrum of the log power spectrum. The power cepstrum of a signal containing an echo exhibits a strong peak at the delay of the echo, even for signals whose autocorrelation function does not have such a peak. To avoid confusion, instead of using the terms frequency, magnitude, and phase, a ripple in the (log) frequency domain is described by its *quefrency*, *gamnitude*, and *saphe*. So the power cepstrum is a function of quefrency, expressed in units which are equivalent to the spatial units of the original signal (e.g., pixels).

Since the power cepstrum was first proposed, several closely related transforms have been defined, which are reviewed here for completeness. The development of homomorphic techniques for deconvolution and separation of multiplied signals [59]

gave rise to the *complex cepstrum*, defined as the inverse z-transform of the complex logarithm of the z-transform of a signal. The real component of the complex cepstrum is the *real cepstrum*, also defined as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of a signal. For a 1-D discrete input signal $c(x)$, these definitions may be stated as

$$\text{Power cepstrum:} \quad \left| \mathcal{F} \left[\log \left| \mathcal{F} \left[c(x) \right] \right|^2 \right] \right|^2 \tag{4.3a}$$

$$\text{Complex cepstrum:} \quad \mathcal{Z}^{-1} \left[\log \left(\mathcal{Z} \left[c(x) \right] \right) \right] \tag{4.3b}$$

$$\text{Real cepstrum:} \quad \mathcal{F}^{-1} \left[\log \left| \mathcal{F} \left[c(x) \right] \right| \right] \tag{4.3c}$$

where $\mathcal{F}$ is the Discrete Fourier Transform (DFT) and $\mathcal{Z}$ is the z-transform, respectively, and the superscript $^{-1}$ indicates inverse transform. In most applications of echo analysis, the power cepstrum is used to identify the echo arrival times, and the complex cepstrum (in which phase information is preserved) is used to recover the underlying waveform.

The definition of the power cepstrum suggests a simple procedure for recovering the monocular disparity over a finite region (referred to as a *window*) of the composite image: take the Fast Fourier Transform (FFT) of the composite image window, compute the squared magnitude of the result and take its log (yielding the log power spectrum), perform a second FFT, take its squared magnitude yielding the power cepstrum, and output the quefrency value of maximum cepstral response (within a range of expected disparity values).

However this procedure ignores the fact that the composite image is purely real, and therefore its (log) power spectrum is even-symmetric (and real). Hence the FFT of the log power spectrum is purely real (and even-symmetric). Why take the squared magnitude of a signal that is purely real? In doing so, sign information is lost. This sign indicates the "phase" or *saphe* of the ripple in the log spectrum: positive for cosine saphe, negative for sine saphe. Substituting for $H_d(\omega)$ in Eqn. (4.2) from

Eqn. (3.8b), the log power spectrum of the composite image may be written as

$$\log |C(\omega)|^2 = \log(2 + 2\cos \omega d) \; + \; \log |S(\omega)|^2 \tag{4.4}$$

Therefore the ripple in the log power spectrum arising from the echo process is in cosine saphe, and the corresponding peak in the FFT of the log power spectrum (at quefrency d) is positive, whereas other peaks (at quefrencies other than d) may be negative. In the power cepstrum, these positive and negative peaks are indistinguishable since both become positive in the operation of taking the squared magnitude.

As an alternative to the power cepstrum, this final step of taking the squared magnitude of the result of the second FFT, can be replaced by taking the *real component* of the result of the second FFT. In this way, negative peaks (corresponding to ripples in sine saphe) can be ignored as noise in searching for the correct monocular disparity. The real component of the FFT of the log power spectrum (normalized by the number of sample points in the input signal) is henceforth referred to as simply "the cepstrum", symbolically represented as

$$\mathcal{K}\left[c(x)\right] = \frac{1}{N} Re\left\{ \mathcal{F}\left[\log |\mathcal{F}\left[c(x)\right]|^2 \right] \right\} \tag{4.5}$$

where $\mathcal{F}$ is the DFT (of which the FFT is an implementation) and N is the number of sample points in the input signal $c(x)$. This is the operation used in this thesis to estimate the visual echo delay in a composite image window.

As an example of echo detection by the cepstrum, the same artificially generated composite signals analyzed by autocorrelation in Sec. 3.3.3 (see Fig. 3.7) were analyzed by the cepstrum. Like autocorrelation, the cepstrum was successful in detecting the correct echo delay for a white noise single signal (Fig. 4.1a,b). Unlike autocorrelation, as the single signal was low-pass filtered, the cepstrum remained successful in detecting the correct echo delay (Fig. 4.1c-h).

The relationship between autocorrelation and the cepstrum provides some insight as to why the cepstrum is a more effective echo detector. The two are very similar with the exception of the logarithm operation inserted between Fourier transforms in

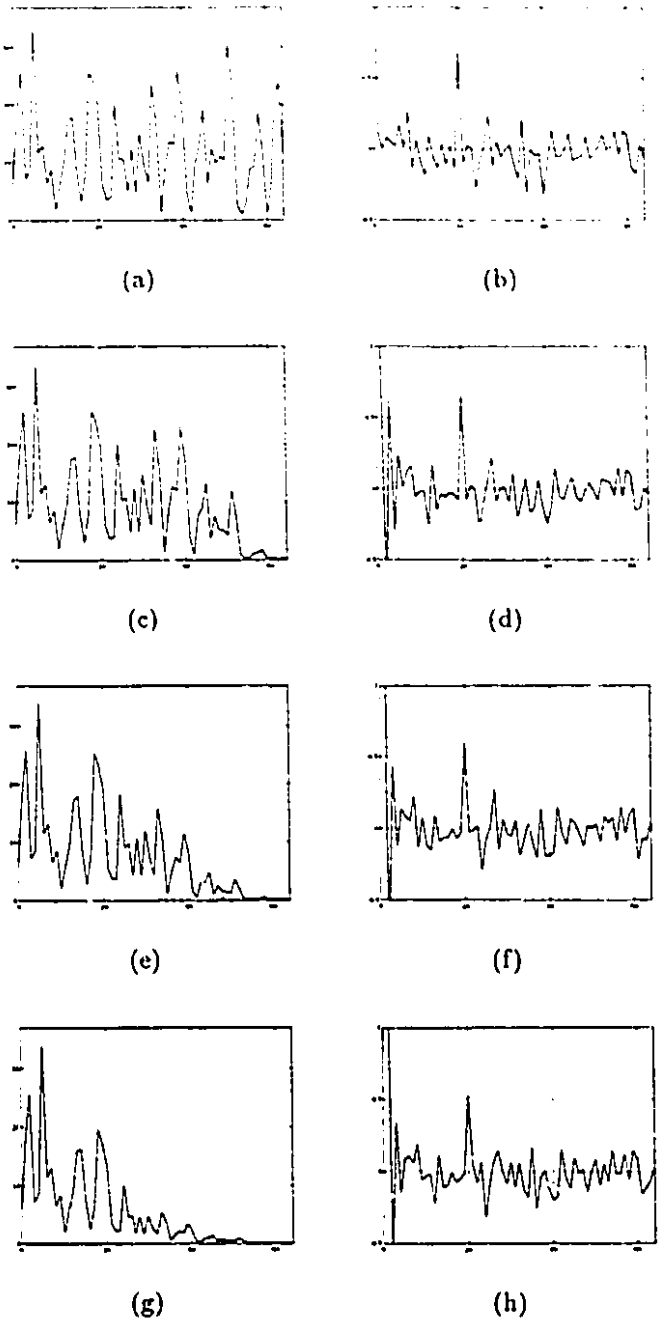


Figure 4.1: Cepstrum as a means of detecting monocular disparity. **(a)** The Fourier spectrum of a composite signal, generated from an original signal consisting of white noise, and echoed with a delay of 20 sample points. **(b)** The cepstrum of the composite signal whose spectrum is given in (a). The echo delay of 20 is successfully detected. **(c,e,g)** The same Fourier spectrum as in (a), low-pass filtered with a decreasing cut-off frequency. **(d,f,h)** Cepstra of the signals whose spectra are given in (c,e,g) respectively. Unlike autocorrelation, the cepstrum remains successful at detecting the echo delay as the original signal deviates from white noise.

the cepstrum. The logarithm is a compressive nonlinearity which reduces the relative dominance of high values over small values. Therefore when the composite image spectrum contains much more energy at low frequencies than high frequencies (as is the case in natural imagery [21]), or contains high spikes at some frequencies, the logarithm tends to make the spectrum more uniform, so that it more closely resembles the ripple of the echo transfer function. In terms of image structure, the logarithm tends to reduce the effect of periodic patterns and slow, smooth intensity variations, all of which interfere with detection of the visual echo.

4.2 Refining the Cepstrum for Visual Echo Analysis

In this section several techniques used to enhance the performance of the cepstrum for echo detection are reviewed, and their appropriateness to the monocular stereopsis problem evaluated. As other authors have noted [17], the performance of techniques in cepstral analysis is highly data dependent, and those that yield improvements in one domain may be detrimental in another. Theoretical or empirical justification is provided as to why a particular tool is or is not applicable to visual echo analysis and the monocular stereopsis problem in particular.

4.2.1 Zero-padding the Composite Signal

It is common when performing frequency analysis of short discrete signals to increase their length by appending zeros to each data window. This increases frequency resolution in the discrete Fourier spectrum (as provided by an FFT operation) at the expense of additional computation. It has been reported that zero-padding of a composite signal improves echo detection by the cepstrum [44, 17]. This is attributed to the increased "sampling rate" of the composite power spectrum, which reduces aliasing in the cepstrum. When applied to visual echo analysis, zero-padding of the composite image window is most effective when the image data is forced to have a mean value of zero. This is easily accomplished by subtracting from each intensity value in the image window, the mean value of intensities in the window. Without

the zero-meaning operation, leakage of the zero-frequency (DC) value becomes visible in the composite power spectrum due to the increased sampling rate. Since in natural imagery the DC value is very high relative to the rest of the spectrum [21], the sinc-like ripple due to leakage of the DC value tends to obscure the ripple due to the visual echo. By forcing the DC value to zero (by zero-meaning the composite signal) this problem is avoided, and zero-padding has a beneficial effect on the performance of echo detection. In theory, the more zeros appended to the composite signal, the better the performance of the cepstrum. In practice, there is a limit to which it is worth paying for the extra computation. Once the composite image sequence has been zero-padded to a length of 2048 points, further zero-padding incurs large computational costs, for only a marginal improvement in performance.

4.2.2 Improving Computational Efficiency

In terms of computational complexity, the cepstrum is dominated by two N-point FFTs, where N is the length of the input sequence, requiring $O(N \log N)$ operations each. However, since the input to both FFTs is a purely real sequence, the Hartley transform may be used to compute the same result with better efficiency [71]. The discrete Hartley transform of a sequence $y(x)$ is defined as

$$Y(k) = \sum_{x=0}^{N-1} y(x) \left(\cos \frac{2\pi kx}{N} + \sin \frac{2\pi kx}{N} \right) \tag{4.6}$$

and unlike the Fourier transform, involves no complex arithmetic. The even-symmetric component of the Hartley transform (HT) of a signal is equal to the real component of the Fourier transform (FT) of the signal. The odd-symmetric component of the HT is equal to the negative imaginary component of the FT. The Fast Hartley Transform (FHT) has the same computational complexity as the FFT algorithm, but in practice requires approximately 50% less data memory and 40% less execution time [71].

The FHT may be substituted for each FFT in the computation of the cepstrum as follows. The first FFT is used to compute the power spectrum of the composite image sequence. If $C(k)$ is the Hartley transform of the composite image sequence,

this power spectrum is given by $[C(k)]^2 + [C(N-k)]^2$ [15]. The second FFT is used to compute the Fourier transform of the log power spectrum — a real, even sequence. The FT and HT of a real, even signal are identical, so in this case the FHT may be substituted directly for the FFT.

4.2.3 Ineffectiveness of Windowing and Smoothing

Another technique commonly used in conjunction with FFT operations is to apply a non-rectangular window function (e.g., Hanning, Hamming, or Blackman window) to the input data sequence. In the frequency domain, these functions have lower side lobes than the sinc function corresponding to a rectangular window, thereby reducing leakage in the output of the FFT. Unfortunately such windowing of a composite image sequence has a negative effect on echo detection by the cepstrum [17]. The use of a window function which is not constant over its entire length is equivalent to distorting the original signal relative to its echo. In other words, windowing of the composite image sequence is inconsistent with the visual echo.

It has also been suggested that echo detection by the cepstrum in the presence of additive noise is improved by windowing the log spectrum [33]. Windowing the log spectrum is equivalent to smoothing the cepstrum, which may in fact smooth out the peak due to the visual echo. Given the relatively low levels of additive noise generally present in a composite image, and the desire for maximum resolution in the cepstrum for the purposes of sub-pixel monocular disparity measurement, such smoothing of the cepstrum is undesirable.

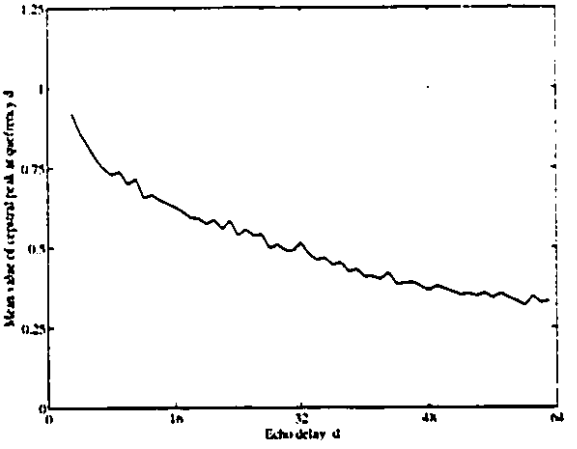
Other authors have reported that Hanning smoothing (convolution with [0.25 0.5 0.25]) of the log spectrum improves echo detection by the cepstrum [44]. Smoothing of the log spectrum is equivalent to windowing of the cepstrum. Assuming there is some *a priori* range of monocular disparity values, it is preferable to search for the highest peak over some interval of the cepstrum rather than modify the entire sequence. However, caution is needed in selecting this disparity range. The cepstrum will always exhibit a high peak at zero quefrency, corresponding to the single (unechoed) signal. For natural images (and any other non-white signal), the first

few values of the cepstrum after zero quefrency will also be relatively high, due to correlation between neighbouring pixels in the single image. This suggests that the disparity search range should be limited to quefrencies greater than some minimum value (denoted by τ_c), determined empirically from the class of images under study. This value constitutes a lower bound on the range of measurable disparity values. The upper bound is given by half the length of the composite image sequence input to the cepstrum (since the cepstrum is an even symmetric function).

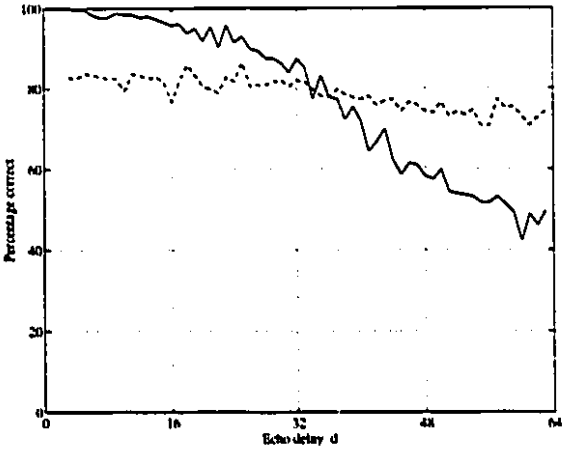
4.2.4 Echo Truncation and Bias in the Cepstrum

The height of the peak in the cepstrum at quefrency corresponding to the echo delay d, is a crucial factor in cepstral analysis. The greater the height of this peak, the greater the likelihood that it is the maximum value of the cepstrum over a given search interval. This peak height is influenced by a number of factors: the single image power spectrum $|S(\omega)|^2$, the relative magnitude of the single image and its echo (in the case of identical apertures in monocular stereopsis, unity), and the degree of overlap between the single image and its echo. Since the cepstrum is computed over a finite window size, as the delay of the echo increases there are fewer points of overlap. In effect, the echo before the beginning and beyond the end of the window is truncated. Due to this fact the cepstrum is slightly biased toward smaller estimates of the visual echo delay, that is, the larger the echo delay d, the smaller the cepstral peak at quefrency d. The same biasing occurs to the correlation function for finite sequences, and is overcome by *unbiasing* or scaling the raw correlation sequence $\rho(k)$ by $N/(N-k)$, where N is the window size. Due to the nonlinear logarithm operator, the function required to unbias the cepstrum is not as simple.

An experiment was performed to study the implications of echo truncation and how it may be overcome. A natural image was selected for study, horizontally shifted an amount d and added to the original image. The result simulates a composite image with visual echo delay d. From this composite image, 256 arbitrarily chosen 128-point 1–D image windows were extracted. The cepstrum of each image window was computed. The value of the cepstrum at quefrency d, and the quefrency $\hat{d}$ with



(a)



(b)

Figure 4.2: Change in cepstral peak height with increasing echo delay. A natural image was artificially echoed by known quantities and then analyzed by the cepstrum in order to study the behaviour of the cepstral peak as the echo delay (and degree of echo truncation) was increased. Using a 128-point window, 256 scanlines were tested for each echo delay d ranging from 3 to 63. **(a)** The mean height of the cepstral peak at quefrency corresponding to the echo delay d. **(b)** The percentage of trials where the cepstrum successfully identified the echo delay, that is, the cepstral value at d was the maximum value of the cepstrum over the quefrency range $[3, 63]$. The solid line indicates performance of the normal (biased) cepstrum; the dashed line indicates performance of the unbiased cepstrum, where each cepstrum was scaled by the inverse of the curve in (a) so that the height of the cepstral peak at d was unity regardless of d.

maximum value over the quefrency range [3, 63], were recorded. Cepstra where $\hat{d} = d$ were labelled "correct". This procedure was repeated for values of d ranging from 3 to 63. The results indicate that as the echo delay d is increased, the height of the peak at quefrency d decreases nonlinearly (Fig. 4.2a), and performance of the cepstrum deteriorates (solid line in Fig. 4.2b). These results seem to be consistent with previous work suggesting that the cepstral peak becomes submerged in noise with echo truncation greater than 20% [44].

The entire experiment was then repeated but with each cepstrum scaled by the inverse of the curve in Fig. 4.2a, so that the expected height of the cepstral peak at d was one, regardless of the value of d. This unbiasing technique improved performance for larger echo delays, but worsened performance for smaller delays (the dashed line in Fig. 4.2b), the transition occurring at roughly 1/4 the window size. Compared to the biased cepstrum, for large delays the unbiasing increases the height of the correct peak relative to noise, but for smaller delays, the unbiasing emphasizes high quefrency noise relative to the correct peak. Therefore it is concluded that instead of unbiasing the cepstrum, a window size at least four times the maximum expected monocular disparity value should be used, so that echo peaks occur in the range where the biased cepstrum is superior in performance to the unbiased version. Another reason for using a minimum window length of four times the maximum expected disparity will emerge in Sec. 4.4.

4.3 A Model of the Composite Image Cepstrum

Having addressed the issues involved in the computation of the cepstrum, attention is now turned to modelling the form of the composite image cepstrum, and how to best exploit this model in order to measure monocular disparity.

Substituting the expression for the log power spectrum in Eqn. (4.4), into the definition of the cepstrum in Eqn. (4.5), the cepstrum of an N-point sequence (excluding zero-meaning and zero-padding) from a composite image containing an echo of delay

d may be written as

$$\begin{aligned}\mathcal{K}\left[c(x)\right] &= \frac{1}{N}\,Re\left\{\mathcal{F}\left[\log\left(2+2\cos\omega d\right) + \log\left|S(\omega)\right|^2\right]\right\} \\ &= \frac{1}{N}\,Re\left\{\mathcal{F}\left[\log\left(2+2\cos\omega d\right)\right]\right\} + \mathcal{K}\left[s(x)\right] \qquad (4.7)\end{aligned}$$

where $\mathcal{F}$ denotes Discrete Fourier Transform (DFT) and $\mathcal{K}$ denotes the cepstrum defined in Eqn. (4.5).

Using the log series expansion, for an infinite length cosinusoid it is possible to show that [17, 81]

$$\mathcal{F}_\infty\left[\log\left(2+2\cos\omega d\right)\right] = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\,\delta\left(\tau - nd\right) \qquad (4.8)$$

where $\mathcal{F}_\infty$ is the continuous Fourier transform, and τ is the quefrency variable. In the discrete domain, the DFT of $\log\left(2+2\cos\omega d\right)$ is a real-valued, even-symmetric, N-point sequence of alternating-sign "peaks" of decaying height, located at integer multiples of d. When normalized by $1/N$, this result is equivalent to the first term of Eqn. (4.7).

Therefore the visual echo is indicated in the cepstrum not only by a positive peak at quefrency d, but also by a negative peak at $2d$, a positive peak at $3d$, and so on (the peaks at d and $2d$ are referred to as the primary and secondary peaks, respectively). According to Eqn. (4.8), the height of these peaks decays as $1, -0.5, 0.25, \ldots$. In practice, the cepstrum is computed from a finite length, discrete sequence, in which the echo is truncated at the beginning and end of the sequence. Because of this, the observed peaks tend to be smaller in height than what the theory suggests. The experiment described in Sec. 4.2.4 (in particular, Fig. 4.2a) predicted the height of the primary peak in the cepstrum due to a visual echo of delay d. A similar experiment was performed for the secondary peak. The resulting data provides a lookup table for the expected primary and secondary peak heights (denoted by h_1 and h_2) for any echo with delay d relative to the window size N. Notice that h_1 and h_2 are not necessarily equal to the values of the composite image cepstrum at quefrencies d and $2d$, for these

values include the cepstrum of the single image, the second term of Eqn. (4.7).

When d is an integer, the primary peak in the cepstrum occurs exactly at the sample of the cepstrum at quefrency d. In monocular stereopsis, this is highly unlikely. The true displacement between images on the sensor plane arising from two apertures will involve some sub-pixel component. The height of the cepstral peak at $\lfloor d \rfloor$ or $\lceil d \rceil$ will vary according to how far the actual d is from an integer value. One way to model this behaviour is to consider a discrete version of an impulse (a rectangular box, one pixel wide, with unknown height) centered on the true sub-pixel disparity, convolved with a sampling function that integrates over one pixel (a rectangular box, one pixel wide, with height one) [58]. The result is a triangle of width two pixels at the base, sides of equal slope, centered at the sub-pixel disparity.

An experiment was performed to test this model. A simulated composite image was created by adding together two ray traced images of a scene rendered from slightly different viewpoints. The scene consisted of a vertically inclined plane, therefore monocular disparities in the composite image varied smoothly from top to bottom (6.5 pixels to 7.5 pixels), and disparity was constant within each image scanline. Since the scene was artificially generated, these monocular disparity values were precisely known. The cepstrum of each scanline was computed, and samples of the cepstrum at quefrencies 6,7,8 recorded. This data was grouped into bins according to the distance (in quefrency) of each sample from the actual sub-pixel disparity, and the mean cepstral response of the points in each bin computed. The results confirm the triangular peak model proposed in [58] (see Fig. 4.3).

Similar triangular-shaped peaks occur at quefrencies $2d$, $3d$, ..., as predicted by Eqn. (4.8). The cepstrum of the echo impulse response (the first term of Eqn. (4.7)) is given by sampling the resulting waveform at integer locations, so that each triangular peak is represented by two successive samples of the cepstrum.

The cepstrum of the single aperture image (the second term of Eqn. (4.7)) acts as noise in the estimation of the visual echo delay. This function is similar to the spatial autocorrelation function, which for natural imagery has a characteristic shape. Starting at unity for zero lag, the normalized autocorrelation function falls off rapidly

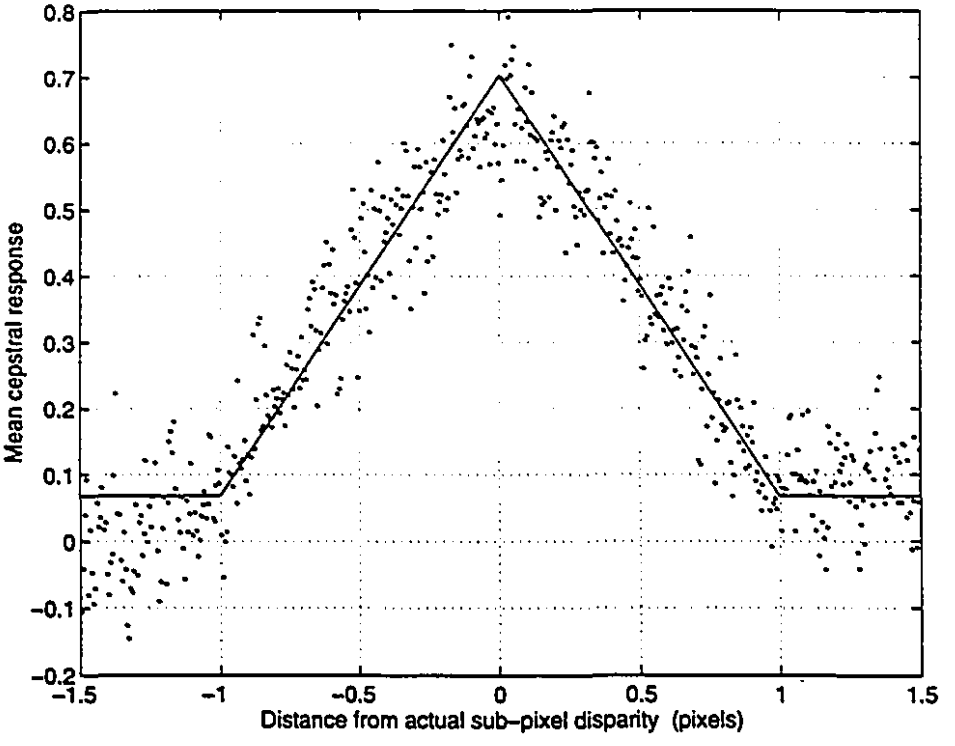


Figure 4.3: Change in cepstral peak height with varying sub-pixel disparity. A simulated composite image was created with monocular disparity varying linearly from 6.5 at the top to 7.5 at the bottom. Horizontal windows of length 128 points were analyzed and values of the cepstrum at quefrencies 6,7,8 recorded. This data was grouped into 500 bins (of approx. 130 points each) according to distance (in quefrency) of each sample from the actual disparity value. The points plotted represent the mean cepstral response in each bin. The solid line indicates the best-fitting triangular peak of width 2 pixels at the base.

to some lag τ_c, from which point on it is roughly constant, asymptotically approaching the ratio of the square of the mean to the mean squared intensity value [70]. The single image cepstrum is expected to exhibit similar behaviour, with τ_c depending on the size and density of texture elements in the image. However, if the single image contains a spatially periodic texture, strong peaks may occur in the single image cepstrum at quefrencies corresponding to the period of the texture (and its harmonics). This poses a problem for monocular stereopsis and will likely result in incorrect disparity estimates (periodic textures pose a similar problem for binocular stereopsis).

To further examine the nature of the single image cepstrum, a study was performed on two (single) natural images. Overlapping 1-D image windows of length 128 points were extracted from the two images and their cepstra computed, for a total of 65,536 single image cepstra. In the first experiment, the pointwise mean and standard deviation of all of these cepstra were calculated (see Fig. 4.4a). The mean has a high value at zero quefrency (approx. 9), then falls off rapidly to a value near zero. For quefrencies $\tau > \tau_c$, where $\tau_c = 3$, the mean cepstral value is roughly constant, near zero. The standard deviation of the single image cepstra exhibits similar behaviour, with a slight peak at zero quefrency, then roughly constant for all other quefrencies.

To what extent are neighbouring samples of the cepstrum correlated? In the next experiment, the mean of the autocorrelation functions of all the single image cepstra was computed (see Fig. 4.4b). For non-zero lags, the mean autocorrelation function is roughly zero, suggesting that the single image cepstrum can be approximated as an uncorrelated sequence (i.e., white noise). Finally, what is the distribution of values of the single image cepstrum? Separate histograms of the cepstral values at quefrencies 12,24,36,48,60 pixels were generated from the single image cepstra (see Fig. 4.4c). The five histograms are all fairly similar; the differences between them reflect the structure of the two natural images. For example, one of the images contains a tablecloth with a periodic pattern. Since this pattern resembles an echo, cepstra in this region of the image have high values at quefrency corresponding to the period of this apparent

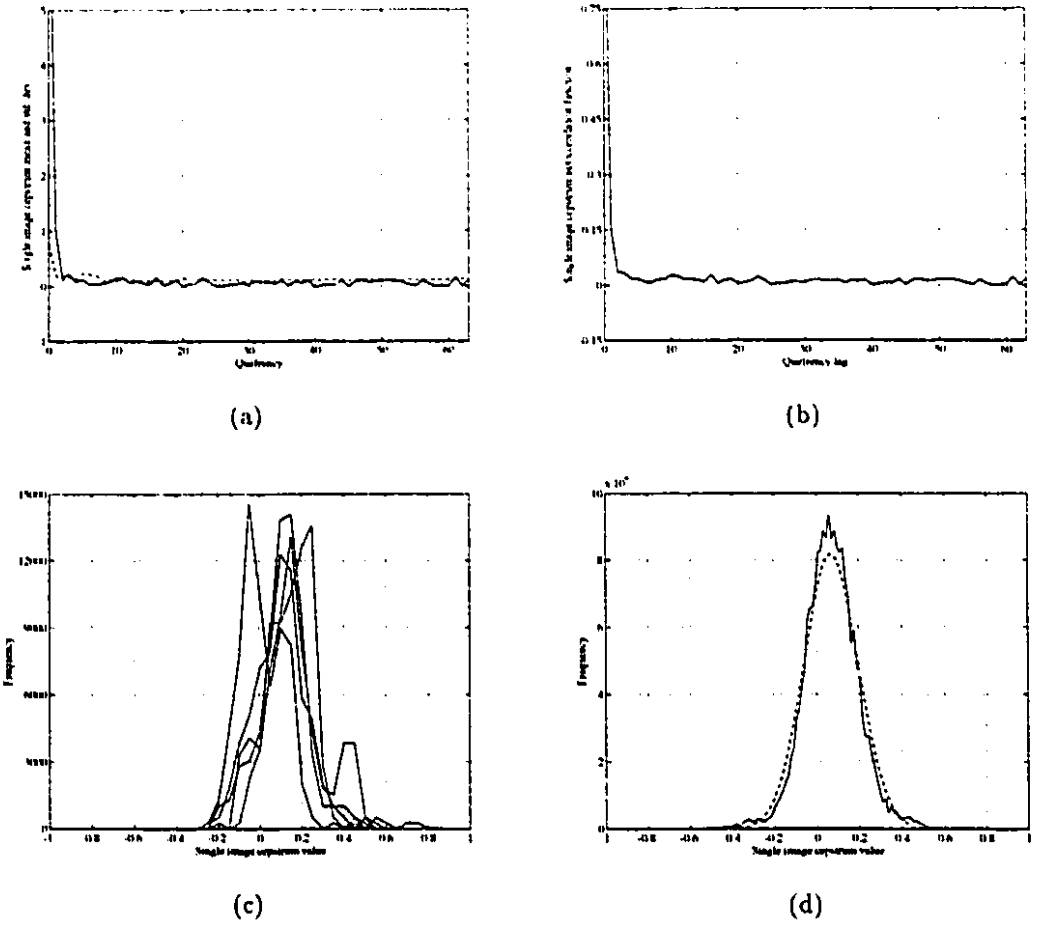


Figure 4.4: Statistical behaviour of the single image cepstrum. In this experiment, the cepstra of 65,536 128-point windows taken from two single natural images were computed. **(a)** The mean (solid line) and standard deviation (dashed line) of the single image cepstra at each quefrency. **(b)** The mean autocorrelation function of the single image cepstra. **(c)** Histograms of single image cepstral values at quefrencies 12,24,36,48,60. **(d)** A histogram of all cepstral values at quefrencies within [3,63]. A Gaussian distribution given by the mean and standard deviation of this pooled data set is shown as a dashed line for comparison.

echo. This explains why one of the histograms has a secondary mode at a cepstral value of approx. 0.4. Despite the small differences between these histograms, the first and second order statistics of the single image cepstra do not change significantly across quefrency (see Fig. 4.4a).

Therefore, for quefrencies $\tau > \tau_c$, the single image cepstrum can be modelled as a stationary sequence. In other words, it is assumed that the distribution of single image cepstrum values is the same at all quefrencies $\tau > \tau_c$. With this assumption, all the single image cepstra data over the quefrency range [3, 63] can be pooled to form an estimate of this stationary distribution (see Fig. 4.4d). A Gaussian distribution with mean and variance given by this data is superimposed on the histogram as a dashed curve. It is clear from the result that the single image cepstrum values for quefrencies $\tau > \tau_c$, can be well-modelled as Gaussian distributed.

This now completes a model of the composite image cepstrum. The model consists of a waveform of triangular peaks two pixels wide, with height $h_1, h_2, ...$, centered at quefrencies $d, 2d, ...$, and sampled at integer locations (top curve of Fig. 4.5). Added to these sampled peaks is a stationary Gaussian white noise sequence, with mean μ_s and variance σ_s^2 (bottom curve of Fig. 4.5). This model forms the basis for reliable estimation of the monocular disparity value as described in the next section, and the derivation of a confidence measure associated with this estimate, described in Sec. 4.6.

4.4 Measuring Monocular Disparity from the Cepstrum

One way to estimate monocular disparity from the cepstrum is to simply find the quefrency, over the range of expected disparities, with maximum cepstral value. In light of the model of the cepstrum introduced in the last section, there are two areas in which this technique may be improved. First, the pattern of repeating triangular peaks can be exploited to help select the correct peak due to the echo, and second, disparity can be measured to sub-pixel precision.

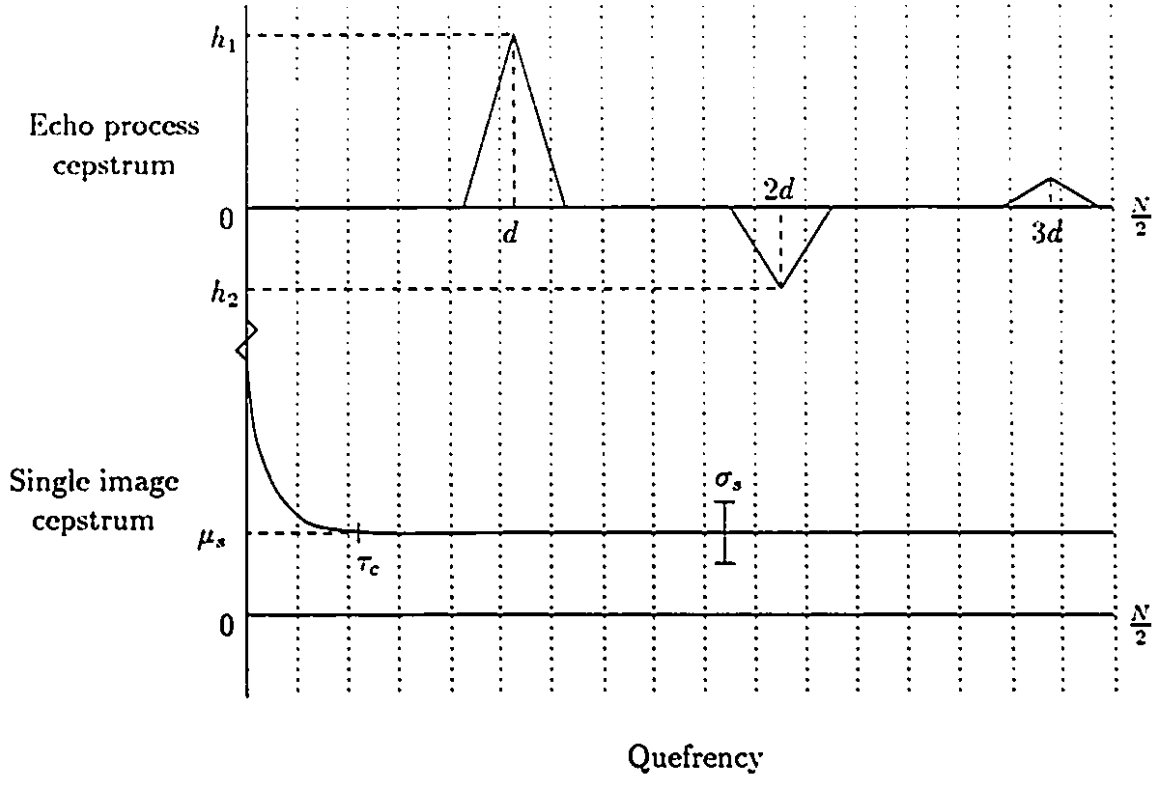


Figure 4.5: Components of the composite image cepstrum. The top curve represents a model of the cepstrum of the visual echo process, consisting of triangular peaks of height $h_1, h_2, \ldots$ at quefrencies $d, 2d, \ldots$ where d is the spatial echo delay or monocular disparity. The bottom curve represents a model of the single image cepstrum, which for quefrencies $\tau > \tau_c$ is modelled as an uncorrelated, stationary Gaussian process with mean μ_s and variance σ_s^2. The composite image cepstrum is given by the sum of these two waveforms, sampled at integral quefrencies (represented by the dotted lines).

4.4.1 Selecting the Correct Peak

if a cepstral peak at quefrency τ is accompanied by a negative peak at 2τ (and if the window size is sufficient, a positive peak at 3τ), this provides further evidence for τ being the correct disparity value. This evidence may be accumulated at the site of the original peak, τ [20]. For example, if $\kappa(\tau)$ is the composite image cepstrum, one may form a modified cepstral sequence

$$\kappa'(\tau) = \kappa(\tau) - \kappa(2\tau) + \kappa(3\tau) \tag{4.9}$$

and identify the quefrency with maximum $\kappa'(\tau)$ as the delay of the visual echo. However, based on the model introduced in the last section, some important observations can be made regarding this technique. The single image cepstrum, modelled as independent Gaussian noise with variance σ_s^2, is present in every value of $\kappa(\tau)$. Therefore in Eqn. (4.9), the variance of noise in $\kappa'(\tau)$ is $3\sigma_s^2$. However, the peaks due to the visual echo, at quefrencies $d, 2d, 3d, \ldots$, are *not* independent. If the primary peak is weaker than expected, the secondary and tertiary peaks are also weaker than expected, often completely obscured by noise. Therefore, in practice, this technique for accumulating evidence from multiple peaks may act to reinforce noise peaks and suppress the correct peak, leading to more incorrect disparity estimates than if the primary peak were used alone. For this reason this technique is not recommended.

Another feature of the composite image cepstrum model is the triangular shape of the function relating the value of the cepstrum to distance from the true sub-pixel disparity. According to this model, a difficulty arises when the true disparity is roughly 0.5 pixels from its nearest integer. In this case, the two cepstral values on the triangle centered at d will both be approximately half the value of the "true" peak at d. Therefore it is more likely that some noise peak in the cepstrum will be higher than the peak due to the visual echo, resulting in an incorrect disparity estimate. A simple solution to this problem is to interpolate the value of the true sub-pixel cepstral peak at every successive pair of points in the cepstrum, and select the disparity with maximum interpolated peak value. Assuming the model of the

cepstrum in the last section, if the true disparity, d, lies in the interval $[\tau, \tau+1]$, the expected value of the cepstrum at d is the height of the apex of the triangular peak, or

$$\kappa(\tau^*) = \kappa(\tau) + \kappa(\tau+1) - \mu_s \tag{4.10a}$$

where

$$\tau^* = \tau + \frac{\kappa(\tau+1) - \mu_s}{\kappa(\tau) + \kappa(\tau+1) - 2\mu_s} \tag{4.10b}$$

is the quefrency at which the apex occurs, the *expected* value of d.

Eqn. (4.10a) gives the interpolated peak height between any two successive points in the cepstrum. The maximum interpolated peak height is a much better technique for peak selection, compared to simply taking the maximum sample of the cepstrum. Since μ_s is constant over the interval of the cepstrum under consideration, it may be removed from this computation. This leads to a simple and elegant solution to the problem of incorrect peak selection due to non-integer echo delays: select the *maximum pairwise sum* of the cepstrum as the peak due to the visual echo.

4.4.2 Sub-pixel Disparity Localization

The composite image cepstrum is a discrete signal, yet monocular disparity varies continuously with depth in the scene. As first described in Sec. 3.1, the relationship between depth and monocular disparity is nonlinear. At some depths, small differences in disparity correspond to large differences in depth; at other depths, the reverse is true. For example, suppose some application is concerned with measuring depth between 0.5 and 5 m using the double aperture CCD camera described in Sec. 6.1. Assume the camera is focused at a depth of infinity (see the solid-line curve in Fig. 3.2). In this example, an error in disparity of 0.5 pixels corresponds to an error in depth as great as 1.2 m, 27% of the operating range. Therefore in order to discriminate significant differences in depth, it is necessary to estimate disparity to sub-pixel precision.

Assuming the model of a triangular peak two pixels wide, a simple way to obtain sub-pixel disparity estimates is to interpolate the location of the apex of this triangle

[58], as given by Eqn. (4.10b). However this technique is incomplete, for it ignores the single image cepstrum, that acts as Gaussian white noise added to each point of the composite image cepstrum. This noise tends to perturb the triangular peak shape. Interpolating sub-pixel disparity based solely on two points of the cepstrum is analogous to fitting a straight line to noisy data by connecting two points. Although this scheme is sufficient for a rough estimation of interpolated peak height (as is used in selecting the correct peak), it is not sufficient for precise estimation of sub-pixel disparity.

As an alternative, based on the more complete model of the composite image cepstrum (see Fig. 4.5), it is possible to develop a maximum likelihood (ML) estimate of sub-pixel disparity. Rather than use two points of the cepstrum to interpolate a peak, this method seeks a disparity value that best accounts for the *entire* observed cepstrum. In a sense, the technique fits a function to the observed cepstrum. This function is similar to the model of the composite image cepstrum developed in Sec. 4.3, but is simplified to have only one one variable parameter — the monocular disparity value d. This parameter is chosen to minimize the sum of squared errors between the function and the observed cepstrum. To continue the analogy of fitting a straight line to noisy data, this technique is analogous to the familiar linear regression method.

The maximum likelihood estimate of monocular disparity is developed as follows. First, assume the maximum pairwise sum of the cepstrum correctly identifies the neighbourhood of the true disparity value. Let $\dot{d}$ have the higher cepstral value of these two points, $\dot{d}+1$ the other (a similar result is obtained in the opposite case). Based on the triangular peak model, this suggests the true disparity d lies in the interval $[\dot{d},\ \dot{d}+0.5]$. However, this triangle is perturbed by noise arising from the single image cepstrum. Therefore this interval is extended by half a pixel on either side, so that the true disparity is assumed to lie in the interval $[\dot{d}-0.5,\ \dot{d}+1]$.

The function that is fit to the observed cepstrum to estimate this disparity is given by the expected value of the composite image cepstrum developed in Sec. 4.3, which

may be written as

$$f_d(\tau) = \begin{cases} \mu_s - h_1(|\tau - d| - 1), & \text{for } d-1 < \tau < d+1, \\ \mu_s - h_2(|\tau - 2d| - 1), & \text{for } 2d-1 < \tau < 2d+1, \\ \mu_s, & \text{otherwise} \end{cases} \tag{4.11}$$

for quefrencies $\tau > \tau_c$, and where the echo peaks at quefrencies nd, $n > 2$, are assumed to be negligible. In general, the parameters h_1 and h_2 vary according to d, but over the interval $[\bar{d} - 0.5,\ \bar{d} + 1]$ they are assumed to be constant. The value of these constants can be determined from the lookup table of expected primary and secondary peak heights (described in Sec. 4.3). The parameter μ_s can be determined by removing quefrencies less than τ_c, and the primary and secondary peaks from the observed cepstrum, and taking the mean of the remaining samples. This leaves one undetermined model parameter: the monocular disparity value d. The maximum likelihood choice of d is that which maximizes the probability of obtaining the observed cepstrum, assuming that $f_d(\tau)$ is the "true" cepstrum. In what way does the observed cepstrum differ from $f_d(\tau)$? According to the composite image cepstrum model, the observed cepstrum is given by $f_d(\tau)$ plus Gaussian white noise of variance σ_s^2. In the presence of Gaussian noise, the maximum likelihood criterion reduces to choosing d to minimize

$$e_d = \sum_{\tau=\tau_c}^{N/2} [\kappa(\tau) - f_d(\tau)]^2 \tag{4.12}$$

where $\kappa(\tau)$ is the observed composite image cepstrum. Note that the choice of d has already been limited to the interval $[\bar{d} - 0.5,\ \bar{d} + 1]$, since it was assumed that the peak selection process correctly identified the peak in the cepstrum due to the visual echo.

The regression function $f_d(\tau)$ is more complicated than a line. It consists of piecewise linear segments, the parameter d determining where these segments begin and end. Because of this, the error function e_d is not a well-behaved function (see Fig. 4.6). To facilitate the task of minimizing Eqn. (4.12), this error function may be broken into 0.5 pixel wide subintervals that *are* well-behaved, and can therefore be

differentiated. Since d has already been limited to the interval $[d-0.5,\ d+1]$, only three such subintervals need be considered: $[\bar{d}-0.5, \bar{d}]$, $[\bar{d}, \bar{d}+0.5]$, and $[\bar{d}+0.5, \bar{d}+1]$. In what follows, one of these subintervals will be examined in detail; the other two are developed in a similar manner.

Over the subinterval $[\bar{d}, \bar{d}+0.5]$, Eqn. (4.11) becomes

$$f_{\bar{d}}(\tau) = \begin{cases} \mu_s - h_1(|\tau - \bar{d}| - 1), & \tau = \bar{d}, \bar{d}+1 \\ \mu_s - h_2(|\tau - 2\bar{d}| - 1), & \tau = 2\bar{d}, 2\bar{d}+1 \\ \mu_s, & \text{otherwise} \end{cases} \tag{4.13}$$

Substituting this expression into Eqn. (4.12) gives

$$c_{\bar{d}} = \sum_{\tau=\bar{d}}^{\bar{d}+1} [\kappa(\tau) - \mu_s + h_1(|\tau - \bar{d}| - 1)]^2 + \sum_{\tau=2\bar{d}}^{2\bar{d}+1} [\kappa(\tau) - \mu_s + h_2(|\tau - 2\bar{d}| - 1)]^2 + \beta \tag{4.14}$$

where β is a constant, the squared difference between the observed cepstrum and the mean μ_s over the range $[\tau_c, N/2]$, excluding $\bar{d}, \bar{d}+1, 2\bar{d}, 2\bar{d}+1$. Differentiating Eqn. (4.14) with respect to d, setting the derivative equal to zero and solving for d gives the ML disparity estimate for this subinterval,

$$d^*_{[\bar{d},\bar{d}+0.5]} = \frac{h_1^2(2\bar{d}+1) + 2h_2^2(4\bar{d}+1) + h_1[\kappa(\bar{d}+1) - \kappa(\bar{d})] + 2h_2[\kappa(2\bar{d}+1) - \kappa(2\bar{d})]}{2(h_1^2 + 4h_2^2)} \tag{4.15}$$

If d^* lies outside the subinterval $[\bar{d}, \bar{d}+0.5]$, then $c_{\bar{d}}$ has no local minimum in the subinterval $[\bar{d}, \bar{d}+0.5]$. In other words, the ML disparity estimate lies in another subinterval.

For the other two subintervals the expression for d^* is as follows:

$$d^*_{[\bar{d}-0.5,\bar{d}]} = \frac{h_1^2(2\bar{d}-1) + 2h_2^2(4\bar{d}-1) + h_1[\kappa(\bar{d}) - \kappa(\bar{d}-1)] + 2h_2[\kappa(2\bar{d}) - \kappa(2\bar{d}-1)]}{2(h_1^2 + 4h_2^2)}$$

$$d^*_{[\bar{d}+0.5,\bar{d}+1]} = \frac{h_1^2(2\bar{d}+1) + 2h_2^2(4\bar{d}+3) + h_1[\kappa(\bar{d}+1) - \kappa(\bar{d})] + 2h_2[\kappa(2\bar{d}+2) - \kappa(2\bar{d}+1)]}{2(h_1^2 + 4h_2^2)}$$

It is possible that the ML disparity estimate may lie at the junction between

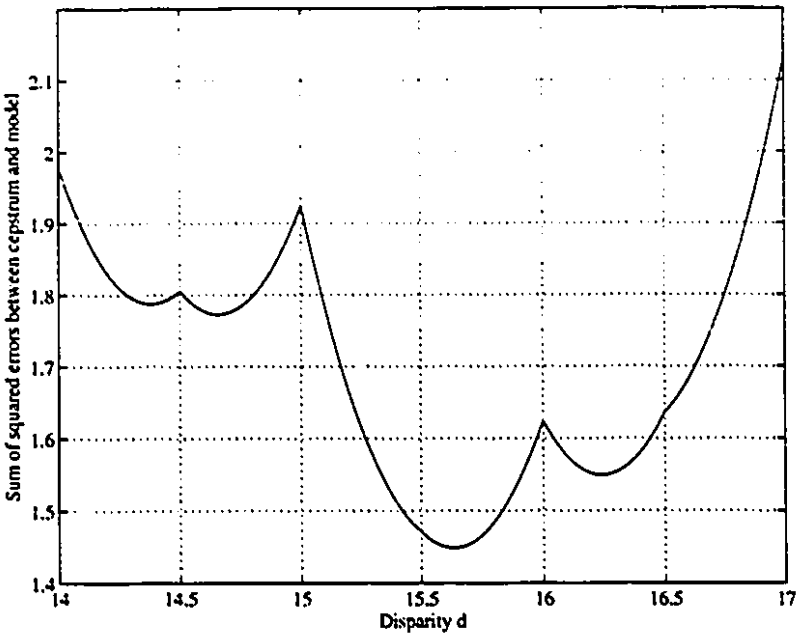


Figure 4.6: Sum of squared errors between observed cepstrum and regression function. An example of the function e_d in Eqn. (4.12), the squared error between an observed composite image cepstrum, and the function to be fit to the cepstrum, given by Eqn. (4.11). In this case, the true disparity lies between 15.5 and 16. Note that over 0.5 pixel intervals, the function is differentiable.

two subintervals. To allow for this possibility, if e_d has no local minimum in the subinterval under consideration, the global minimum (i.e., one of the two endpoints) is returned as the best estimate for that subinterval. The estimate from the three subintervals with minimum error is output as the estimated sub-pixel disparity.

Performance of this technique, compared to the simple interpolation scheme suggested by the triangular peak model [58] was evaluated on a simulated composite image with known sub-pixel disparity values. The creation of this composite image is described in Sec. 4.3, in the description of Fig. 4.3. The differences between the estimated and actual disparity values for each composite image window were recorded for three different techniques of disparity estimation. The resulting histograms of disparity error indicate the superiority of the ML disparity estimate over other techniques. It should also be noted that computation of the ML estimate requires the evaluation of three simple expressions (as in Eqn. (4.15)) and several comparisons, only marginally more computation than the other techniques.

4.5 Effects of Blur and Noise

The model of the composite image cepstrum is based on an idealized model of the composite image in which there is no out-of-focus blur or camera noise, as in Eqn. (3.5). A real composite image acquired by a double-aperture camera will not be so ideal, and is better described by Eqn. (3.12). Therefore it is important to understand the impact of noise and blur on the cepstrum, and to evaluate the technique for monocular disparity measurement in the presence of noise and/or blur. For example, if the technique is very sensitive to noise, then high quality (and more expensive) optics and image acquisition hardware may be required for the range sensor. If blur is a major problem, smaller apertures may have to be used, requiring greater scene illumination. On the other hand, if the technique is relatively insensitive to noise and/or blur, the range sensor can be constructed of less expensive components, without affecting performance.

Performance of the cepstrum for detecting an echo in noise has been a concern

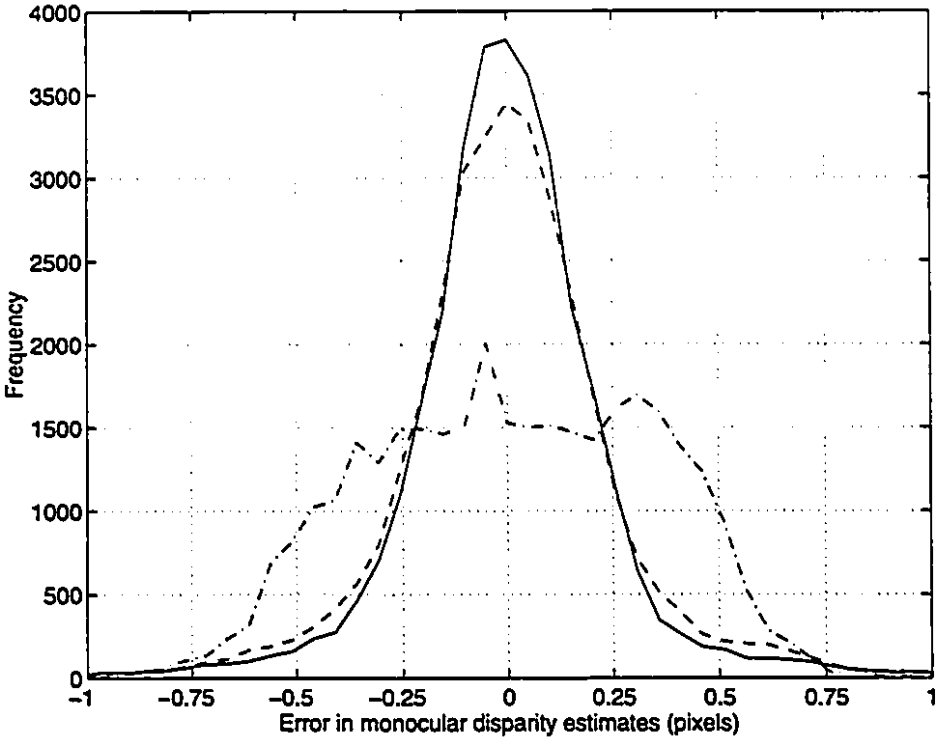


Figure 4.7: Evaluation of sub-pixel disparity localization techniques. A simulated composite image with known sub-pixel disparity values was analyzed by the cepstrum with a 128-point window. Estimated disparities were compared to ground truth and a histogram of error values computed. The dash-dot line is for the best integer estimates, the error roughly uniformly distributed between -0.5 and 0.5. The dashed line is for the simple interpolation scheme suggested by Eqn. (4.10b). The solid line, having the best performance, is for the maximum likelihood sub-pixel disparity estimates.

almost as long as the cepstrum has been used for echo analysis [13, 44]. Gaussian white noise added to the single signal and its echo is represented in the cepstrum by two effects: a reduction in the height of the peak at the echo delay, and the addition to the entire cepstrum of an extra noise field [32]. The degree of these effects is determined not only by the signal-to-noise ratio (SNR), but also by the relative bandwidth of signal and noise. In the case of monocular stereopsis, the signal is a natural image (which tends to have little energy at high frequencies, i.e., narrowband [21]), and the noise is modelled as being white (containing roughly equal energies at all frequencies, i.e., broadband). Since the noise bandwidth is significantly greater than the signal bandwidth, noise is more of a concern in monocular stereopsis than it may be in other domains.

To measure the effect of noise on overall performance, an experiment was performed with an artificially generated composite image where the actual disparity values were precisely known (the formation of this image is described in Sec. 4.3). Windows of length 128 points were extracted from the composite image and the monocular disparity estimated using the technique in Sec. 4.4. To reduce the effect of gross disparity errors, instead of reporting the mean or root-mean-squared error, the 90th percentile absolute error value was computed. That is, 90% of disparity estimates over the composite image had an absolute error less than this reported value. The experiment was repeated as increasing levels of Gaussian white noise were added to the composite image, from a signal-to-noise (SNR) ratio of 80 dB (almost no noise) to 0 dB (extreme noise). The results are presented in Fig. 4.8. Noise has very little effect on monocular disparity estimates for an SNR greater than 30 dB. Given the relatively high SNR of current CCD cameras (50–60 dB), it is concluded that camera noise *alone* is not a significant factor in determining performance of this range sensor.

Out-of-focus blur is a more significant problem for monocular stereopsis, since the depth cues of the visual echo and blur circle diameter covary with depth in the scene (assuming the two camera apertures are not ideal pinholes). From Eqn. (3.12), it is apparent that the blurring kernel, $B_a(x,y)$, acts in the same manner as the echo impulse response, $h_d(x,y)$, that is, it is convolved with the single aperture image.

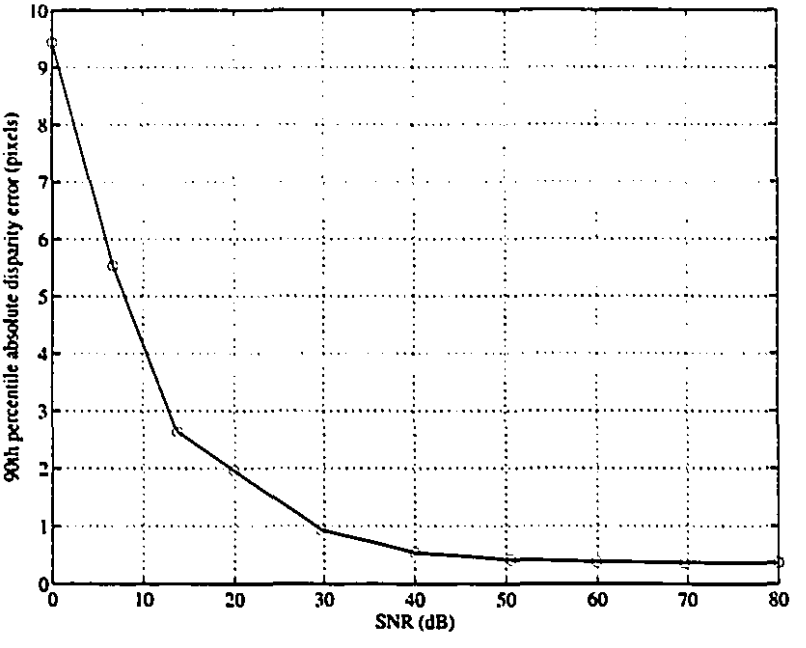


Figure 4.8: Effect of camera noise on monocular disparity detection. A simulated composite image with known sub-pixel disparity values was analyzed by the cepstrum under increasing levels of artificially generated Gaussian white noise. A window size of 128 points, a disparity range of 5-20 and the maximum likelihood method was used to estimate sub-pixel disparity throughout the image. For each level of noise tested, the absolute disparity error at the 90th percentile (i.e., 90 % of pixels have an estimated disparity closer to ground truth than this value) is plotted.

Following the operations of the cepstrum, the first FT transforms this convolution into multiplication, the logarithm transforms multiplication into addition, and the second FT is linear. Therefore the cepstrum of the convolution of two signals is the sum of their individual cepstra. So the effect of blurring a composite image with the kernel $B_a(x, y)$, is to add the cepstrum of $B_a(x, y)$ to the composite image cepstrum.

What is the cepstrum of the blurring kernel? In 1-D, the cepstrum of the pillbox blurring kernel of Eqn. (3.10) consists of a low quefrency hump, followed by negative spikes at integer multiples of the pillbox diameter (see Fig. 4.9a,b). The cepstrum of the 1-D Gaussian blurring kernel with $r = 1/2$ in Eqn. (3.11) has a larger low quefrency hump and falls off like an exponentially decaying sinusoid (see Fig. 4.9c,d). If the disparity search range includes the low quefrencies most affected by out-of-focus blur, one can expect a large number of incorrect disparity estimates in the presence of either form of blur. However, if disparities in the scene are relatively large compared to blur circle diameter, one would expect blur to have little impact on the performance of disparity measurement. At high quefrencies the cepstrum of the blurring kernel is essentially zero, so high quefrency peaks in the composite image cepstrum are undisturbed by the addition of the cepstrum of the blurring kernel.

In practice this apparent immunity of the cepstrum to the effects of blur deteriorates, for the following reasons. First, it is inherent to the preceding analysis that blurring is applied to each composite image window via *circular* convolution, that is, the end of each composite image window is blurred into the beginning, and vice versa. In reality, the entire composite image is blurred, and windows are extracted for analysis from the blurred image. Second, in the frequency domain, blurring acts as a low pass filter, strongly attenuating power at all but the lowest frequencies. Therefore compared to the unblurred case, the power spectrum of the blurred composite image will be much smaller in magnitude. This can lead to numerical instability in the computation of the cepstrum. Finally, blur further reduces the signal bandwidth, so that the effect of broadband additive noise is exacerbated, as described above. Due to these reasons, out-of-focus blur has a significant impact on the performance of the monocular disparity estimation by the cepstrum, testimony to the difference between

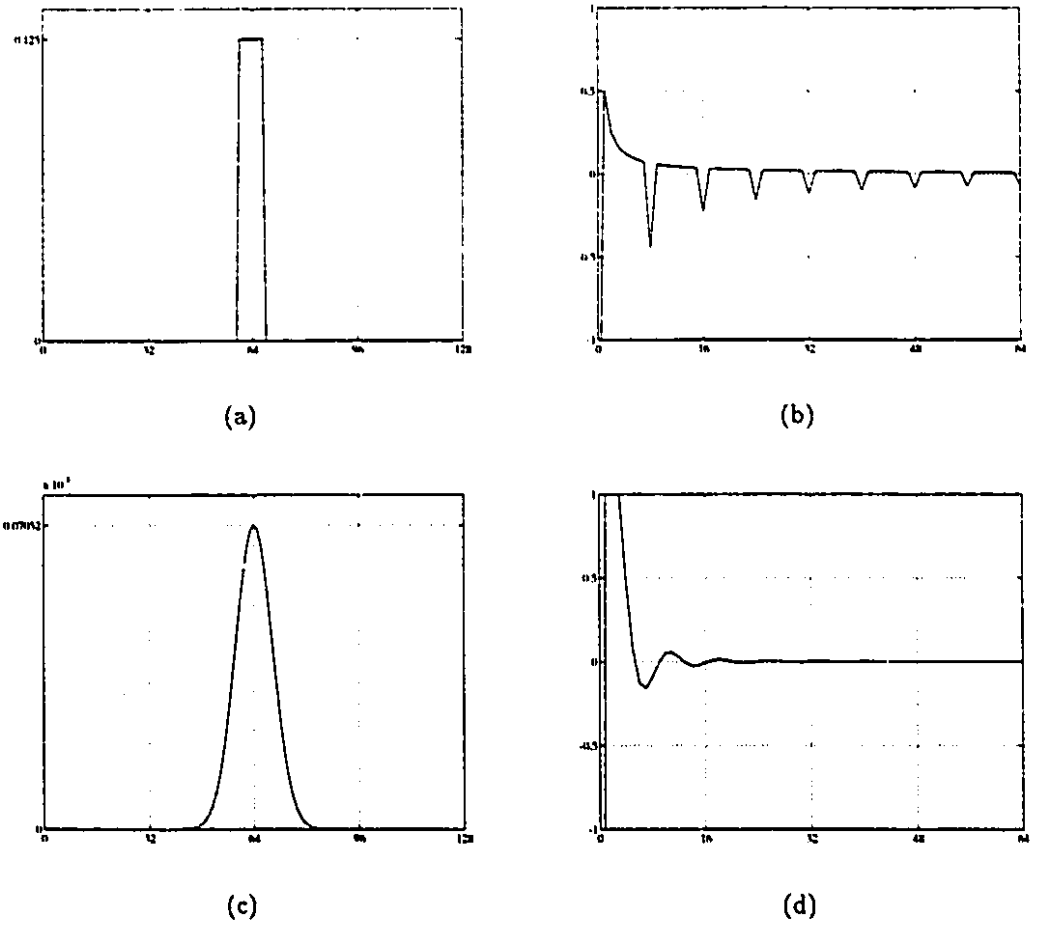


Figure 4.9: Cepstra of two blurring kernels. **(a)** A 1–D pillbox with diameter $a = 8$. **(b)** The cepstrum of (a), consisting of a low quefrency hump and negative spikes at integer multiples of a. **(c)** A 1–D Gaussian with standard deviation $\sigma = \sqrt{r}\,a$ where $a = 8$ and $r = 1/2$. **(d)** The cepstrum of (c), resembling an exponentially decaying sinusoid, approximately zero for quefrencies greater than 12.

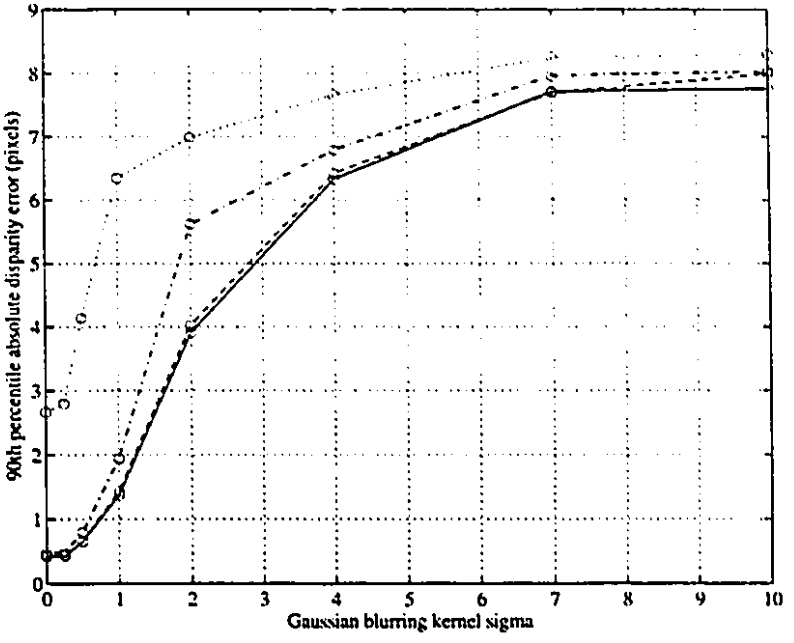


Figure 4.10: Effect of out-of-focus blur and noise on monocular disparity detection. A simulated composite image with known sub-pixel disparity values was analyzed by the cepstrum under increasing levels of Gaussian blurring, with four different levels of additive Gaussian white noise. A window size of 128 points, a disparity range of 15-30 and the maximum likelihood method was used to estimate sub-pixel disparity throughout the image. For each level of blur and noise tested, the absolute disparity error at the 90th percentile (i.e., 90 % of pixels have an estimated disparity closer to ground truth than this value) is plotted. The SNR levels tested were 100dB (solid line), 74dB (dashed line), 47dB (dash-dot line) and 20dB (dotted line).

theory and practice in computation.

To study these effects quantitatively, the experiment used above to examine the effects of noise alone, was repeated for four different SNR levels, under increasing levels of blur. To simulate out-of-focus blur, a 2-D Gaussian of varying width was convolved with the composite image prior to the addition of white noise. Disparities in the scene were set large enough to prevent the echo peak from being drowned in the low quefrency hump of the cepstrum of the blurring kernel. Nonetheless, blur had a major effect on disparity estimates (see Fig. 4.10), even with virtually no additive noise. Notice that at an SNR of 47dB (an average quality CCD camera with current technology), there is little deterioration in performance due to noise alone (Fig. 4.8), but when combined with even a small amount of blur, there is significant deterioration in performance (Fig. 4.10). This provides the motivation for using smaller or slit-shaped apertures to reduce the degree of blur in the horizontal direction, as described in Sec. 3.1.

In the experiment described above, the *unblurred* composite image contained power at relatively high frequencies. This power was removed as the degree of blur was increased. What if the composite image contained very little high frequency power to begin with? Scenes containing regions of roughly constant intensity, or slow, smooth contrast variations, pose a problem similar to out-of-focus blur: insufficient power across the composite image spectrum to identify the ripple due to the visual echo. For this reason, the nature of the composite image texture is a crucial factor in determining the success of monocular stereopsis (as is binocular stereopsis). In general, the more uniform the single image spectrum, the less it will interfere with echo detection. Therefore the best image texture is white noise, containing roughly equal power at all frequencies. The worst image texture is a region of constant intensity, containing power only at zero frequency.

4.6 Confidence Measures

Given an estimate of monocular disparity provided by the cepstrum, what degree of confidence or certainty is associated with it? It is useful in range sensing systems to have a confidence value available for each range estimate. In conjunction with some *a priori* model or assumptions about the scene, this confidence value can be incorporated into surface fitting schemes (e.g., [29, 11, 79]). Most of these surface fitting techniques assume that range estimates are corrupted by additive, Gaussian distributed noise. The estimated variance of this noise then comprises the confidence value. For data provided by monocular (or binocular) stereopsis, this model of errors in disparity estimates is often inappropriate.

In the estimation of monocular disparity from the cepstrum, there are two distinct types of error, which follow naturally from algorithm described in Sec. 4.4. The first kind of error is caused by failure of the maximum pairwise sum of the cepstrum to identify the peak due to the visual echo. If a peak selection error is made, the chosen peak may lie *anywhere* within the disparity search range, suggesting disparity estimates are uniformly distributed over $[d_{min}, d_{max}]$. In other words, if the chosen peak is incorrect, no information is provided as to the correct disparity. The second kind of error is associated with imperfect sub-pixel disparity localization. Assuming the peak in the cepstrum due to the visual echo has been correctly identified, there will be some small error associated with the ML sub-pixel disparity estimate. In this case, disparity estimates are roughly Gaussian distributed about the actual disparity value, with some standard deviation σ_e (see Fig. 4.7). Given these two different kinds of errors, a simple confidence measure such as an estimate of noise variance does not reflect the true distribution of errors in monocular disparity estimates.

Instead, for each disparity estimate, the confidence measure should include the parameters of the two error distributions, and some relative indication as to which distribution applies. This relative indicator is given by the probability that the peak selection process was successful. If the correct peak was identified, the disparity error belongs to the Gaussian distribution, otherwise it belongs to the uniform distribution.

4.6.1 Modelling Errors in Peak Selection

When a peak of the cepstrum is selected as being due to the visual echo, what is the probability that this peak is correct? Given the model of the composite image cepstrum in Sec. 4.3, if the selected peak is *not* due to the visual echo, it must be due to the single image cepstrum. The expected height of the primary echo peak is known, and the single image cepstrum values are modelled as Gaussian distributed. Therefore it is possible to estimate the probability that the selected peak is due to the visual echo, rather than the single image cepstrum. This is performed as follows.

Let $\kappa(\tau)$, $\tau = 0, 1, ..., N-1$ be a composite image cepstrum with parameters μ_s, σ_s, h_1, h_2 and d as defined in Sec. 4.3 (in particular, see Fig. 4.5). Define $Y(\tau) = \kappa(\tau) + \kappa(\tau+1)$, as the pairwise moving sum of the cepstrum, and

$$h = \max_{d_{min} \leq \tau < d_{max}} \left\{ Y(\tau) \right\} = Y(\bar{d}) \tag{4.16}$$

as the output of the peak selection process. The goal is to estimate the probability that $\bar{d}$ is correct given h. Here "correct" means that the true disparity d lies in the 1.5 pixel interval (around $\bar{d}$) considered by the sub-pixel disparity localization process. By this definition, if $\bar{d}$ is correct, any discrepancy between the estimated and true disparity is due the sub-pixel disparity localization process alone. Using Bayes Rule the desired probability may be written as

$$P(\,\bar{d} \text{ correct} \mid Y(\bar{d}) = h) = \frac{p_1}{p_1 + p_2} \tag{4.17a}$$

where

$$p_1 = p(\,Y(\bar{d}) = h \mid \bar{d} \text{ correct})\; P(\,\bar{d} \text{ correct}) \tag{4.17b}$$

$$p_2 = p(\,Y(\bar{d}) = h \mid \bar{d} \text{ NOT correct})\; P(\,\bar{d} \text{ NOT correct}) \tag{4.17c}$$

The sequence $Y(\tau)$ is the pairwise moving sum of the cepstrum, which, excluding the peaks due to the visual echo, is Gaussian distributed with mean μ_s and variance σ_s^2, de-

noted by $\sim \mathcal{N}(\mu_s, \sigma_s^2)$. Therefore, excluding the visual echo, $Y(\tau) \sim \mathcal{N}(2\mu_s, 2\sigma_s^2)$. The expected value of $Y(\tau)$ differs from $2\mu_s$ in the neighbourhood of the peaks due to the visual echo. If the true disparity occurs between τ and $\tau + 1$, the expected value of $Y(\tau)$ is $h_1 + 2\mu_s$, where h_1 is the primary peak height given by the lookup table described in Sec. 4.3.

Therefore if $\bar{d}$ is correct, one is tempted to assume that $Y(\bar{d}) \sim \mathcal{N}(h_1 + 2\mu_s, 2\sigma_s^2)$, $Y(2\bar{d}) \sim \mathcal{N}(h_2 + 2\mu_s, 2\sigma_s^2)$, and for all other $\tau \in [d_{min}, d_{max} - 1]$, $Y(\tau) \sim \mathcal{N}(2\mu_s, 2\sigma_s^2)$. This allows analytic derivation of expressions for p_1 and p_2, thus providing the probability that $\bar{d}$ is correct given h. However, several factors make this derivation quite complex. First, note that $Y(\bar{d}) = h$ is the *maximum* of $\{M = d_{max} - d_{min}\}$ random variables, and second, each $Y(\tau)$ is *correlated* with $Y(\tau - 1)$ and $Y(\tau + 1)$, because $Y(\tau)$ is a pairwise moving sum. Therefore the expressions for p_1 and p_2 involve M-dimensional integrals of an M-dimensional Gaussian probability density function (PDF), requiring expensive numerical computation. Furthermore, the assumption stated at the beginning of this paragraph does not hold. If $d - \lfloor d \rfloor > 0.5$, the secondary echo peak occurs at $Y(2\bar{d}+1)$ instead of $Y(2\bar{d})$, and the points of $Y(\tau)$ neighbouring the peak locations do not have expected value $2\mu_s$, since they overlap the triangular peaks at d and $2d$.

Instead of delving into a complex and lengthly probabilistic analysis, it is more practical to estimate the required probability distributions using a Monte Carlo simulation. In terms of evaluating these probabilities, there are only two independent parameters that specify the model to be simulated. They are the disparity range $M = d_{max} - d_{min}$, and the expected height (above $2\mu_s$) of the primary peak in $Y(\tau)$ (normalized by σ_s), denoted by α. Given these two parameters, a Monte Carlo simulation can be performed to precompute a lookup table, which allows the probability that $\bar{d}$ is correct, to be computed from the height of the selected cepstral peak, $h = Y(\bar{d})$. Once this lookup table is generated, it may be used with any monocular disparity estimate provided by the cepstrum of any composite image window.

The simulation to generate the lookup table proceeded as follows. In each trial, a sequence of M samples from a $\mathcal{N}(0,1)$ distribution was generated, representing

the single image cepstrum. A real-valued number d was chosen at random from the interval $[1, M-1]$ to represent the true disparity, and a triangular peak of height α centered at d was sampled at integer locations and added to the cepstrum. Next, the pairwise sum of the cepstrum was formed, and the maximum value $h = Y(\bar{d})$ determined. This process was repeated for 100,000 statistically independent trials. After each trial, the normalized [1] peak height $\bar{h} = (h - \alpha)/\sqrt{2}$ was recorded, along with a bit indicating if the chosen peak location $\bar{d}$ was correct or not (according to the definition above). The resulting data was sorted by peak height and grouped into bins of equal size. In each bin the percentage of trials correct was calculated. These values were rescaled so instead of ranging from chance (the performance of peak selection if it chose $\bar{d}$ at random) to 100%, they range from 0 to 100%. The resulting curve has a familiar shape resembling the cumulative probability distribution function for a Gaussian PDF. Therefore a sigmoid curve of the form

$$p(\bar{h}) = 1 - \frac{1}{2}\operatorname{erfc}(\varepsilon_1 \bar{h} + \varepsilon_2) \tag{4.18}$$

is fit to the points using a nonlinear least-squares technique (e.g., Levenburg-Marquardt [63]), where erfc() is the complementary error function (see Fig. 4.11), and ε_1 and ε_2 determine the horizontal stretch and horizontal shift of the sigmoid, respectively. This entire simulation is repeated for various values of M and α, and the curve parameters $(\varepsilon_1, \varepsilon_2)$ recorded in a two-dimensional lookup table.

The number of trials used in a Monte Carlo simulation should not be an arbitrarily chosen parameter. For a given number of trials and some level of confidence, error bounds may be computed for quantities determined from the results of the simulation [69]. Similar principled techniques are available for selecting the number and size of the bins described above. Although such considerations are important in Monte Carlo simulations for probability density estimation [18], here the simulation is simply being used to generate points for fitting a curve. This curve is not a probability density

[1] $Y(\bar{d}) \sim \mathcal{N}(\alpha, 2)$; to normalize h to a $\mathcal{N}(0,1)$ distribution, subtract the mean and divide by the standard deviation

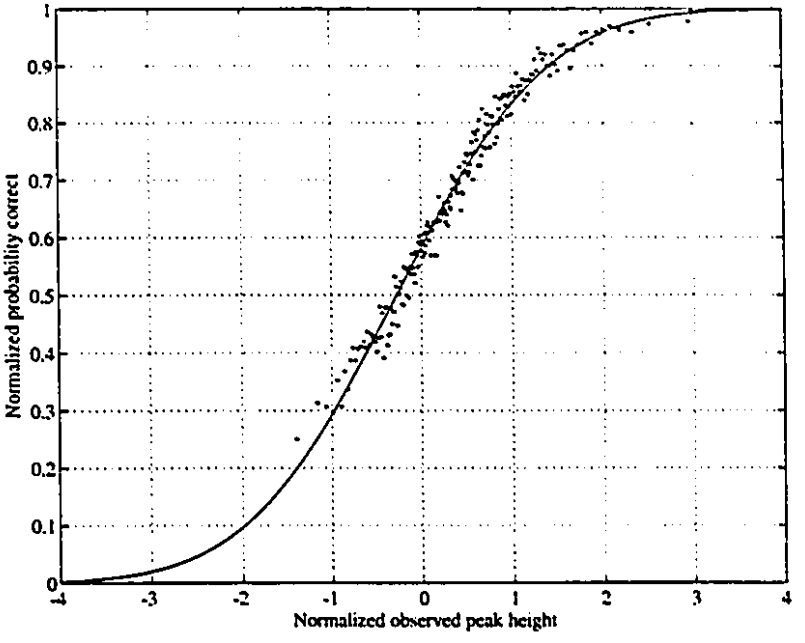


Figure 4.11: Probability correct as a function of normalized peak height. The output of a Monte Carlo simulation with model parameters $M = 20$ and $\alpha = 3$. In this example 100,000 independent trials were performed and the results grouped into 200 bins of 500 points each. The horizontal axis is the normalized peak height, $\tilde{h}$. The solid line indicates a function of the form Eqn. (4.18) fit to the data using a nonlinear least-squares technique.

function, but rather a function that estimates the probability of correct peak selection for a given normalized peak height. The validity of this estimation (and the choice of the simulation parameters that gave rise to it) is evaluated and shown to be accurate by the exercise depicted in Fig. 4.12a.

Once this lookup table is generated, it may be applied to determine the probability of correct peak selection for any cepstrum. The parameters to this lookup table are the disparity range M, which is fixed for a given composite image, and the normalized expected peak height α, computed from a given cepstrum as follows. Having determined the maximum pairwise sum $h = Y(\bar{d})$, the expected primary peak height h_1 for disparity $\bar{d}$ is determined from the lookup table described in Sec. 4.3. The statistics of μ_s, σ_s are calculated from those points of the cepstrum with quefrency greater than τ_c and more than one pixel away from $\bar{d}$ and $2\bar{d}$. The required lookup table parameter is then given by $\alpha = h_1/\sigma_s$. Using bilinear interpolation, the parameters $(\varepsilon_1, \varepsilon_2)$ of the probability distribution associated with M and α are determined from the lookup table. The observed peak height h is normalized to the $\mathcal{N}(0,1)$ distribution, by letting $\bar{h} = (h - h_1 - 2\mu_s)/\sqrt{2}\sigma_s$. This value is substituted into Eqn. (4.18), and the result is scaled back into a value ranging from chance to one (instead of zero to one) yielding the probability that $\bar{d}$ is correct.

To evaluate performance of this confidence measure, the probability correct described above was calculated for simulated cepstra that adhered perfectly to the model of the composite image cepstrum upon which it is based. The resulting data was sorted by probability value and grouped into bins of equal size. In each bin, the actual percent correct was compared to the mean probability value. The result verified that the probability value was being correctly calculated, and that bilinear interpolation from the M,α lookup table was a sufficient approximation (Fig. 4.12a). Probability correct values were then calculated for each disparity estimate in a composite image with known sub-pixel disparities. The results were sorted and binned as described above (see Fig. 4.12b). The probability correct values are generally similar to the actual percent correct values, confirming that this is an effective confidence measure for monocular disparity estimates. The small discrepancy between the ob-

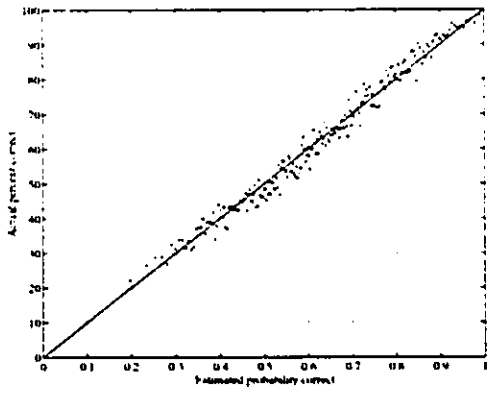
served data points and the ideal line (a line of slope one through the origin) is due to composite image cepstra which do not adhere perfectly to the model. Real single image cepstra may contain outliers caused by underlying periodicity in the single image spectrum, leading to high confidence, yet incorrect, peaks in the composite image cepstrum. This behaviour varies from one composite image to another. For the most part, however, the probability estimates are quite accurate.

It is important to realize that this confidence value is a quantitative probability, not a qualitative, ordinal measure of "degree of confidence". Therefore it may be used directly in Bayesian and other forms of probabilistic analysis. This value can be interpreted as the probability that error in a disparity estimate is due to the sub-pixel localization process alone.

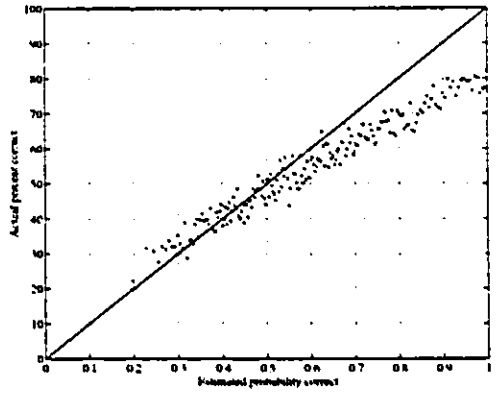
4.6.2 Modelling Errors in Sub-pixel Disparity Localization

If the peak selection process successfully identifies the cepstral peak due to the visual echo, the true disparity lies in the interval of disparities considered by the sub-pixel disparity localization process. The distribution of disparity errors associated with this process may be adequately modelled as Gaussian with mean zero and some standard deviation σ_e. To complete the confidence measure, a reliable estimate of σ_e for each disparity estimate is required.

The sub-pixel disparity localization process is based on finding the disparity over a restricted interval, that minimizes the squared error between the observed cepstrum and a function derived from the model of the cepstrum (see Sec. 4.4.2). Several factors may contribute to errors in the resulting ML disparity estimate. The observed cepstrum may not obey the model in Fig. 4.5, due to aliasing (which may be alleviated by zero-padding), outliers in the single image cepstrum, camera noise, or out-of-focus blur. More significantly, the greater the perturbation of the triangular primary and secondary echo peaks (due to the single image cepstrum), the poorer the ML disparity estimate. This factor, embodied by the measured statistic σ_s, outweighs all others in its effect on σ_e. Furthermore, σ_s captures the effects of aliasing, camera noise, and outliers in the single image cepstrum. Therefore σ_e can be considered to be a function



(a)



(b)

Figure 4.12: Evaluation of probability correct estimates. To evaluate the accuracy of the confidence measure described in this section, probability correct values were calculated for a large number of composite image cepstra with known disparity values. This data was sorted and collected in equal sized bins. For each bin, the mean probability correct was plotted against the actual percent correct. Points along the diagonal line represent perfect probability estimates. (a) If the composite image cepstra adhere perfectly with the model in Sec. 4.3, the probability correct values are essentially perfect. In this example, instead of using a composite image, simulated cepstra were generated according to the model and used as input to the peak selection process. The parameters used here were M=30 and α=3. (b) For an artificial composite image with known ground truth, the estimated probabilities are still quite accurate. This image was analyzed with a window size of 64 points and a disparity range of 10-30, therefore compared to previous examples there are more peak selection errors made. However, for the most part the confidence value correctly identifies those estimates which are incorrect.

of σ_s alone.

The relationship between σ_s and σ_e is a complex one due to the extraneous factors described above and the nature of the sub-pixel disparity localization algorithm. However, the same Monte Carlo simulation used in the previous section can be used here to generate a lookup table which maps observed values of σ_s to expected values of σ_e. Composite image cepstra were generated at random according to the model of Sec. 4.3, and the ML disparity estimate compared to the actual disparity used to generate each cepstrum. This process was repeated for 100,000 statistically independent trials. Cepstra where the peak selection process failed to identify the peak due to the visual echo, were discarded. The standard deviation, σ_e, of errors in disparity estimates from the remaining cepstra was recorded. This procedure was repeated for various levels of standard deviation σ_s in the simulated single image cepstra. The results are plotted in Fig. 4.13. Using simple linear interpolation between points, this curve allows the required value σ_e to be determined from the observed value σ_s.

This completes the confidence measure associated with each monocular disparity estimate. The measure consists of two distributions: a uniform distribution over $[d_{min}, d_{max}]$, and a Gaussian distribution centered on the true disparity with variance σ_e^2. The disparity estimate belongs to the former distribution with probability $1 - q$, and the latter with probability q, where q is the probability that the peak selection process was successful. In Sec. 5.2 it is shown how this confidence measure allows the accurate recovery of 3-D scene structure even if the raw disparity estimates contain many significant errors.

4.7 Summary

This chapter describes an algorithm that takes as its input a window of the composite image and computes its cepstrum. A model of this cepstrum is proposed, that describes how the visual echo manifests itself, and how the single image cepstrum may obscure these cues. Based on this model a two-stage algorithm is given to estimate the monocular disparity value. First, the maximum pairwise sum of samples in the

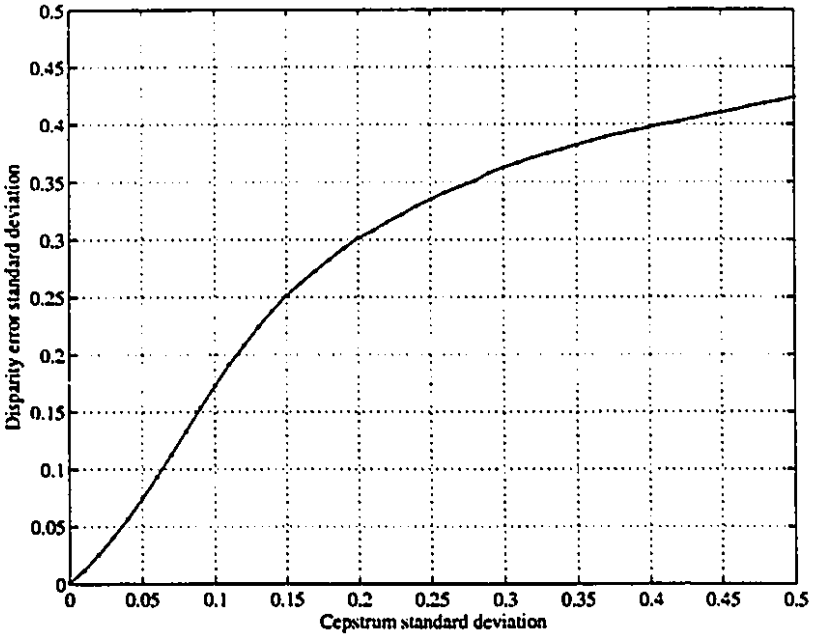


Figure 4.13: Relationship between σ_s and σ_e. A simulation was performed where artificially generated cepstra with different standard deviations σ_s were input to the disparity measurement algorithm. The ML sub-pixel disparity estimate was calculated in a large number of statistically independent trials, and the result compared with the true disparity value. For those cases where the peak selection process was successful, the standard deviation of disparity errors was calculated, giving the σ_e value corresponding to a particular σ_s value.

cepstrum, over the range of expected disparity values, is identified as the peak due to the visual echo. Second, the model of the cepstrum is used to derive a maximum likelihood estimate of the sub-pixel disparity value. This technique is reliable with additive noise of SNR greater than 30 dB, but is sensitive to out-of-focus blur in the direction of the visual echo. A confidence measure is proposed which follows naturally from the algorithm for disparity estimation. Each disparity estimate is modelled to belong to one of two distributions — a uniform distribution over the range of expected disparity values, and a Gaussian distribution centered on the true disparity value. The relative likelihood of these two distributions is given by the probability that the peak selection process correctly identifies the peak in the cepstrum due to the visual echo.

Chapter 5

From Composite Image to Surfaces

The goal of range imaging is to compute the depth of each point in a scene that is visible from a given viewpoint. A composite image acquired by a multiple aperture camera encodes depth by the monocular disparity between image points projected from the same scene point. The last chapter developed a technique to accurately measure the monocular disparity value over a 1-D window of the composite image, and to provide an estimate of the error distribution associated with this measurement. In this chapter, the technique is applied to transform a composite image into an accurate representation of surfaces in the scene. This process consists of two stages. In the first stage, disparity estimates (and the associated confidence measures) are computed at each point on a sampling grid applied to the composite image. The results of this stage are referred to as the *disparity map* and the *confidence maps*. In the second stage, these maps are used in conjunction with some local surface model to compute a piecewise maximum likelihood reconstruction of surfaces in the scene. This technique can generate an accurate approximation of 3-D structure even if the raw disparity map contains many significant errors.

5.1 Computing a Disparity Map

It is inherent to the technique developed in Chapter 4 that estimates of monocular disparity are provided not at a single pixel, but over a region (or window) of the composite image. In the case where disparity is constant over this window, such as in viewing a fronto-parallel plane, the composite image cepstrum obeys the model described in Sec. 4.3, and performance of the algorithm is quite good. The resulting disparity estimate is equally valid throughout the window.

However, in the real world many scenes contain surfaces of varying depth, so that some composite image windows will contain multiple disparities. Even though it was developed for the case of constant disparity, the technique developed in Chapter 4 still provides useful information in the case of non-constant disparity. The problem becomes how to record and interpret this information. Cepstral analysis measures monocular disparity over an entire window, but for the purpose of dealing with disparity variation within a window, this measurement will be recorded *only* at the pixel in the *centre* of the window. By cepstral analysis of windows centered at separate pixels, different disparity measurements can be recorded at separate points within a region of varying disparity. These windows may overlap significantly and still give rise to different disparity estimates. The confidence measure described in Sec. 4.6 indicates which of these estimates are more reliable.

Given this approach, there are two choices to be made in computing a disparity map: the density at which monocular disparity estimates are computed (or equivalently, the degree of overlap between windows for cepstral analysis), and the dimensions of the composite image window upon which these estimates are based.

5.1.1 Disparity Map Density

One strategy to compute a disparity map is to simply divide the composite image into disjoint image windows and compute the monocular disparity for each window. Given a $X \times Y$ composite image and $N \times 1$ window length, the result is a $X/N \times Y$ disparity map. The ratio between the dimensions of the disparity map and the dimensions of

the composite image is referred to as the *density* of the disparity map. In this example, the density of the disparity map is $(1/N, 1)$. Note the anisotropy in density between the horizontal and vertical directions. This is due to horizontally aligned apertures giving rise to a visual echo with a horizontal component only, so that the visual echo may be analyzed independently in each composite image scanline.

Alternatively, windows extracted from the composite image for cepstral analysis need not be disjoint. In fact, they may overlap significantly. Consider a *sliding* composite image window, that is advanced forward by some step size $k \times l$ between the computation of successive cepstra. The density of the resulting disparity map is $(1/k, 1/l)$, regardless of the size of the composite image. In particular, when $k = l = 1$, cepstra are computed for image windows centered on *every* pixel, and the resulting disparity map is the same size as the composite image.

How is the density of the disparity map (as determined by k and l above) to be chosen? In general, the more variation in depth across the composite image, the greater the disparity map density required to accurately reconstruct the scene, and the higher the computational cost associated with computing the disparity map. The choice of disparity map density constrains the *resolution* of the final range image — the lower the density, the lower the resolution. Here resolution refers to the horizontal and vertical directions, not resolution in depth. In some applications, low resolution is sufficient, such as evaluating distance to a fronto-parallel plane. In other applications, higher resolution is required to detect and localize objects in 3-D space, such as in mobile robot obstacle avoidance.

5.1.2 Composite Image Window Dimensions

The choice of the composite image window size used for cepstral analysis is constrained by the range of monocular disparities in the scene. The horizontal dimension of the window must be at least four times the maximum expected disparity value. However, there are several other factors that may influence the choice of window size.

Since cepstral analysis is based on estimating the local power spectrum of the composite image, the window size must be large enough to provide enough sample

points for this estimate to be reliable. In general, the longer the composite image sequence input to the cepstrum (assuming a constant echo delay throughout), the more accurate the resulting monocular disparity estimate (fewer peak selection errors, and better sub-pixel disparity localization). The more "challenging" the composite image for cepstral analysis, the larger the window size required in order to obtain reliable disparity estimates. Chapter 4 (in particular, Sec. 4.5) described what factors make a composite image more challenging for echo analysis — camera noise, out-of-focus blur, and lack of image texture.

On the other hand, all of this analysis presupposes that monocular disparity is constant over the entire composite image window. In general, the larger the window size, the less likely that this is true. How is monocular disparity detection by the cepstrum affected when the window contains multiple, significantly different disparities? It is difficult to predict which (if any) of these disparities will be detected. In some cases, the surface with higher spatial frequency texture dominates the cepstrum. In others, the disparity present in the largest portion of the image window is detected. It is certainly not the case that the measured disparity is simply the average disparity over the window. With one constant disparity value over the window, there is one "strong" echo; with multiple disparities over the window, there are multiple "weaker" echoes. The strength of an echo is indicated by the cepstral power at a quefrency corresponding to the delay of the echo. Given the technique for disparity estimation in Sec. 4.4 and the confidence measure in Sec. 4.6, this suggests that when there are multiple disparities over a composite image window (e.g., the window overlaps a depth discontinuity), the probability of correct peak selection is reduced. Therefore these disparity estimates will be associated with much lower confidence values compared to the ideal, constant disparity case.

This presents somewhat of a dilemma. In order to obtain reliable disparity estimates in the presence of noise, blur, or poor texture, a large window size is required. However, if this window size is too large, there may be many significantly different disparities over the window, leading to a less reliable disparity estimate. One solution to this problem is to provide more samples for the cepstrum by extending the window

in the vertical dimension, while leaving the horizontal dimension at the minimum required value.

In computer vision, it is often assumed that the world is composed of piecewise continuous surfaces. Under this assumption, depth varies slowly over most of a range image, and therefore monocular disparity varies slowly over most of a composite image. Therefore at a given horizontal position, the monocular disparity in one scanline should be very similar, if not the same, as in neighbouring scanlines. The image data in neighbouring scanlines can provide additional sample points for the cepstrum, with less (compared to increasing the length of a 1-D window) risk of encountering significantly different disparities over the window.

Since the orientation of the visual echo is known (i.e., the orientation of the two apertures relative to the image plane), there is no need to compute a 2-D cepstrum of the composite image. Doing so also introduces additional computational expense. Instead, given a 2-D image window, successive scanlines can be concatenated, forming a long 1-D sequence. Assuming the disparity value is the same or very close in successive scanlines, this is similar (in terms of performance of the cepstrum) to using a long 1-D composite image window. There are two differences: the image texture in successive scanlines is likely to be more similar than over an extended 1-D region, and in the concatenated sequence, there are interruptions in the visual echo at points where successive scanlines are concatenated. This interruption is the same kind that occurs at the beginning and end of a composite image window due to echo truncation (see Sec. 4.2.4). One possible solution to this problem is to concatenate the scanlines in different orders, compute the cepstrum for each concatenation order, and take the mean of these cepstra. The peak due to the true disparity should be present in all the cepstra, while any artifacts in the cepstrum introduced by concatenation will vary with concatenation order, and thus partially cancel each other.

5.2 Surface Reconstruction

After completing cepstral analysis of windows extracted from the composite image, the output consists of the following information at each point on an image sampling grid: (1) a monocular disparity estimate with sub-pixel precision, and (2) a confidence measure for this estimate, consisting of (a) the probability that the correct cepstral peak was selected, (b) the standard deviation of error in the sub-pixel disparity estimate, and (c) the prior disparity range $d_{max} - d_{min}$. This section describes a method for interpreting data in this format to compute an explicit representation of surfaces in the scene, a process often referred to as surface reconstruction.

If the density of the disparity map is less than (1, 1), then there are pixels in the composite image for which there are no disparity estimates, and hence no depth can be computed. Surface reconstruction provides a representation of the scene that is not attached to any discrete grid. This representation provides depth at *any* point in the scene, not only at positions where disparity was calculated, nor only at positions defined by the pixel grid of the composite image. The disparity map itself may contain errors, both large errors due to incorrect peak selection, and small errors due to noise in the sub-pixel disparity localization process. The distribution and relative likelihood of these two types of errors is given by the confidence measure. This confidence measure is not particularly useful alone; it should be combined with the disparity map to form a better approximation of the scene than that provided by the disparity map itself. Finally, in many applications, the 3-D structure of the world can be described in terms of some high level model. The goal of the range imaging process is the determine the parameters of this model for an observed scene, based on raw data from the range sensor. The surface reconstruction method described in this section provides a framework for this process.

As described in Sec. 2.5, there are a number of difficulties with surface reconstruction techniques that seek a globally optimal solution surface, such as the thin plate or thin membrane energy models [29, 79, 11]. The approach taken here is quite different. The composite image is divided into regions, where within each region, it

is assumed that 3-D structure in the scene can be described by some local surface model. These regions can be analyzed independently (i.e., in parallel) to determine the "best fitting" local surface model for each region. The problem is to define what best fitting means in terms of the data provided by cepstral analysis, and to select an appropriate local surface model.

Unlike the so-called *robust* methods often used in computer vision [53, 67, 73, 72], the technique developed in the next section does not explicitly assume that the input data consists of two distinct classes, genuine data and outliers. Instead, a model is fit to the data which maximizes the likelihood of the disparity estimates and confidence values provided by cepstral analysis. As opposed to a binary classification of the input data, this technique uses a continuously varying probability to indicate degree of certainty. The choice of an appropriate class of models is determined by the application domain in which the range sensor is being used. The more sophisticated the *a priori* knowledge of the environment, the more sophisticated this model may be.

5.2.1 A Maximum Likelihood Framework

Rather than convert disparity values into absolute depth and perform surface reconstruction in the depth domain, there are advantages to performing surface reconstruction in the monocular disparity domain. If surfaces were reconstructed in 3-D space, not only would disparities have to be converted to depth, but the estimates of sub-pixel disparity error would also have to be converted. Because the relationship between disparity and depth is nonlinear, the same degree of uncertainty at two different disparities translates into quite different degrees of uncertainty at two different depths. It is preferable that uncertainty be related only to the measurement process, not the value of the measurement itself. A similar problem occurs in the conversion of a uniform distribution of disparities (as occurs when there is a cepstral peak selection error) into the corresponding distribution of depths. The resulting depth distribution is no longer uniform and is significantly more complex. Also, if surface fitting is done in the disparity domain, the surface model itself can be converted from disparity to

depth, rather than each individual disparity estimate. Of course, the relationship between a given surface model in the two domains must be well understood. For example, a plane in depth is a plane in monocular disparity (see Appendix A for proof), whereas higher order models may not share this duality.

Suppose that over some region of the composite image, disparity in the scene can be described by the R-parameter surface model

$$d = D(x, y;\ a_1, a_2, \ldots, a_R)\ . \tag{5.1}$$

Given a set of measurements $\{(x_i, y_j, d_{ij})\}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, the goal is to determine the parameters of the surface model that "best fits" the data. Assume there is no uncertainty in the x_i and y_j values, since they represent column and row position in the composite image. If uncertainty in the disparity values d_{ij} can be modelled as a Gaussian distribution, then the problem reduces to that of least-squares surface fitting. Representing this uncertainty by estimates of the standard deviation of error, σ_{ij}, at each point, the least-squares criterion is to set the model parameters to minimize

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[\frac{d_{ij} - D(x_i, y_j;\ a_1, a_2, \ldots, a_R)}{\sigma_{ij}} \right]^2 \tag{5.2}$$

referred to as the chi-square value. Assuming that errors in the data points are independent, the resulting surface model is a *maximum likelihood (ML) estimate*. The likelihood of a surface model given by the parameters $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_R$ is defined as the probability of obtaining the data set $\{(x_i, y_j, d_{ij})\}$ assuming that the model $d = D(x, y;\ \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_R)$ is completely true. For independent measurement errors this probability is given as

$$l(a_1, a_2, \ldots, a_R) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left[\ p_D(d_{ij} \mid a_1, a_2, \ldots, a_R)\ \Delta d\ \right] \tag{5.3a}$$

where

$$p_D(d_{ij} \mid a_1, a_2, \ldots, a_R) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} \exp\left\{ -\frac{1}{2} \left[\frac{d_{ij} - D(x_i, y_j;\ a_1, a_2, \ldots, a_R)}{\sigma_{ij}} \right]^2 \right\} \tag{5.3b}$$

is the conditional probability density function for d_{ij}. Maximizing Eqn. (5.3a) is equivalent to minimizing its negative logarithm (with the constant term arising from Δd removed).

$$\Theta(a_1, a_2, \ldots, a_R) = -\sum_{i=1}^{I}\sum_{j=1}^{J} \log\left[p_D(d_{ij} \mid a_1, a_2, \ldots, a_R)\right] \tag{5.4}$$

which for a Gaussian error distribution is equivalent to minimizing the chi-square value of Eqn. (5.2). However, the measurement errors provided by technique outlined in Chapter 4 are *not* Gaussian, therefore the least-squares criterion will not yield the maximum likelihood surface model.

Instead, the conditional probability density function for cepstral disparity estimates is given by the confidence measure described in Sec. 4.6. If the peak selection process correctly identifies the cepstral peak due to the visual echo, the disparity estimate is Gaussian distributed with mean given by the true disparity and standard deviation by σ_c. Otherwise, the disparity estimate is uniformly distributed between d_{min} and d_{max}. The relative probability of these two distributions is given by the "probability correct" estimate, denoted here by q_{ij}. Therefore the required density function is

$$\begin{aligned} p_D(d_{ij} \mid a_1, a_2, \ldots, a_R) &= p_D(d_{ij} \mid \text{peak is correct})\ P(\text{peak is correct}) \\ &\quad + p_D(d_{ij} \mid \text{peak is NOT correct})\ P(\text{peak is NOT correct}) \\ &= \frac{q_{ij}}{\sqrt{2\pi}\,\sigma_{c,ij}} \exp\left\{-\frac{1}{2}\left[\frac{d_{ij} - D(x_i, y_j;\ a_1, a_2, \ldots, a_R)}{\sigma_{c,ij}}\right]^2\right\} \\ &\quad + \frac{1 - q_{ij}}{d_{max} - d_{min}}. \end{aligned} \tag{5.5}$$

Substituting Eqn. (5.5) into Eqn. (5.4) gives the negative log likelihood function (less a constant) for some local model of monocular disparity variation. The location of the global minimum of this function in R-dimensional space gives the ML parameter set. Note the implicit assumption that errors in disparity estimates are independent over the composite image region modelled by a single surface.

Unfortunately, minimization of this negative log likelihood function does not lead to a simple criterion like minimize the chi-square. In fact, the function contains multiple local minima, so the minimization problem is non-convex. Therefore conventional multi-dimensional downhill minimization techniques are not guaranteed to yield the true ML solution. To solve this problem, one may resort to more sophisticated (and potentially slower) minimization procedures such as graduated non-convexity [11] or simulated annealing [24]. Alternatively, a convex minimization procedure can be used with a "good first guess", assuming that the function is convex over a significant neighbourhood around the global minimum. If the initial point is within such a neighbourhood, the minimization process will converge to the global minimum, providing the parameters of the maximum likelihood model.

5.2.2 Surface Approximation by Planar Facets

Minimization of the negative log likelihood function for some model tends to become more problematic and computationally expensive as the complexity of the model is increased. The more complex the surface model, the more likely that the negative log likelihood function contains many local minima. For best results, the model chosen to represent the local structure of monocular disparity values should be a simple, low order model with as few parameters as possible.

Many scenes, particularly in artificial environments, contain surfaces that are locally planar (e.g., walls, doors, tables, floors). For a mobile robot, a locally planar representation of surfaces is adequate for tasks such as navigation and obstacle avoidance. A planar surface in 3-D space corresponds to a planar surface in disparity space (see Appendix A). Taken together, these observations suggest that an appropriate model of disparity over a given composite image region is

$$D(x, y;\, a_1, a_2, a_3) \;=\; a_1 x \;+\; a_2 y \;+\; a_3. \tag{5.6}$$

The size of the composite image region which is modelled as a single plane depends on the scale at which surfaces in the scene can be well approximated as planar. This

region must be large enough to obtain a sufficient number of data points upon which to base the fit, but not so large that the true surface over this region deviates from a planar model. To obtain the highest number of data points over the smallest area, the density of the disparity map should be maximized, that is, a density of (1, 1).

The surface reconstruction algorithm proceeds as follows. The composite image is divided into disjoint patches of the chosen size. The maximum likelihood planar surface is then determined independently for each patch. For the minimization procedure, any a convex multi-dimensional minimization method is sufficient, such as the downhill simplex or Powell's method [63]. More important than the particular minimization algorithm is the choice of an initial solution to guide the minimization process.

Unless the initial solution is within the convex neighbourhood surrounding the global minimum, the minimization procedure is not guaranteed to converge to the ML parameter set. One way around this problem is to run the minimization several times with different starting points, let it converge to a solution each time, and choose the solution with the minimum negative log likelihood value. Several heuristics are available for selecting appropriate initial solutions.

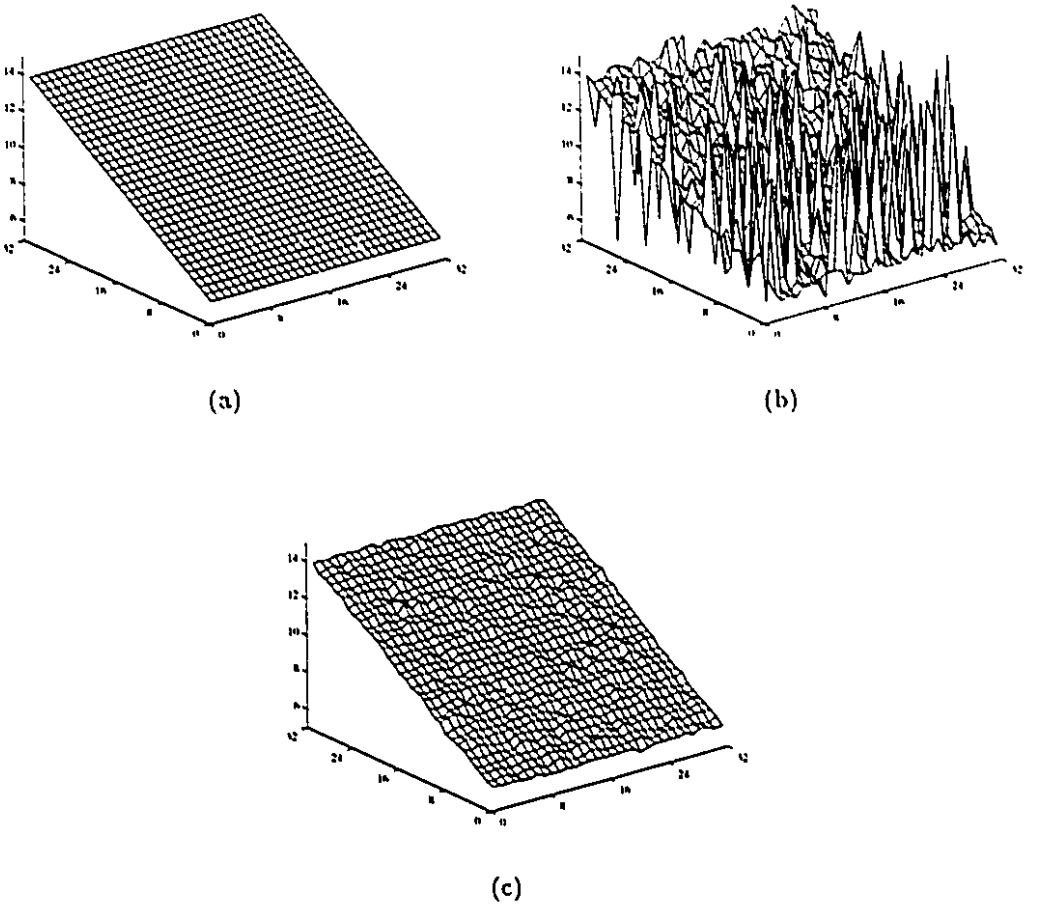
For smooth surfaces, the parameters of adjacent surface patches tend to vary slowly. Therefore an initial solution for one patch may be provided by the final solution from an adjacent patch. This technique may fail in the neighbourhood of depth discontinuities, where the parameters of adjacent surface patches may be significantly different. Since surfaces in range imaging are often fronto-parallel, another possible initial solution is given by a fronto-parallel patch with disparity equal to the median disparity value over the patch. The median is less sensitive to outliers in the disparity map arising from incorrect cepstral peak selection. A third heuristic for selecting an initial solution is to perform a least-squares fit to the highest confidence disparity estimates over the patch. Disparity estimates with high probability correct values are less likely to be outliers, so that a least-squares fit may provide a reasonably good initial solution. Between these three initial solutions, in most cases the minimization process successfully locates the global minimum of the negative log likelihood

function.

As an illustration of this surface reconstruction technique, consider the following simulation. A surface with height ranging from z_{min} to z_{max} was generated over an image grid, and then corrupted by noise in the following manner. To some points, Gaussian noise of standard deviation σ_e was added. Other points were reset to a value chosen randomly between z_{min} and z_{max}. These two events occurred at each point with relative probability q_{ij}, such that globally some specified fraction of the total number of points were in the latter category. The following data was passed from the simulation to the surface reconstruction algorithm described above: σ_e, z_{min}, z_{max}, and at each point on the surface grid: the corrupted height measurement, and the value of q_{ij}. This data simulates the data provided by cepstral analysis of a composite image, with the exception that σ_e does not vary over the surface grid.

The result of applying the maximum likelihood surface reconstruction technique for three different surface classes is presented in Figs. 5.1 through 5.3. In these examples, a 256×256 point surface was corrupted with noise so that 30% of the data points were randomly distributed between 5 and 15, while the remaining 70% were Gaussian distributed about the original true value with $\sigma_e = 0.25$. The resulting surface was reconstructed with maximum likelihood 8×8 planar patches. For each planar patch, initial solutions for the minimization process were given by the ML solution from an adjacent patch, and by a least-squares fit to the 8 highest confidence points within the patch. The original, degraded, and reconstructed surfaces are all displayed as 1/8 resolution mesh plots. The first example (Fig. 5.1) shows the reconstruction of a slanted plane, which can be perfectly modelled by local planar patches. The second example (Fig. 5.2) illustrates that such planar patches can also approximate slowly varying curved surfaces. Finally, an example containing discontinuities (Fig. 5.3) shows that the effect of a discontinuity is limited to those surface patches through which it passes.

In Chapter 6, numerous examples will be given showing the application of this technique to disparity data from real-world composite images. One simple example is presented here to further illustrate the capability of this maximum likelihood surface



(a) (b)

(c)

Figure 5.1: Reconstruction of a planar surface. **(a)** Original surface, a slanted plane with values ranging from 6 to 14. **(b)** Degraded surface, where 30% of points are randomly chosen values between 5 and 15, and the remaining 70% are corrupted by additive Gaussian noise with $\sigma_e = 0.25$. **(c)** Reconstructed surface given by local maximum likelihood planar patches. All surfaces are displayed as 1/8 resolution mesh plots.

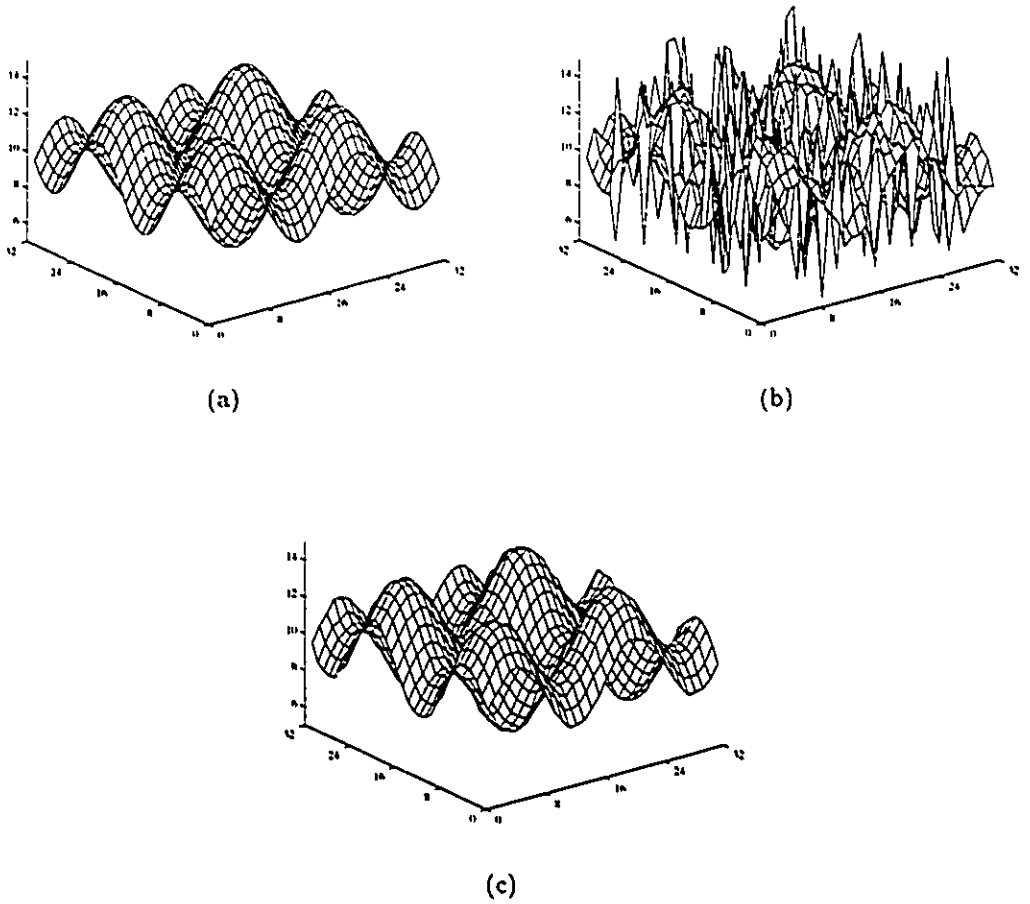


Figure 5.2: Reconstruction of a curved surface. **(a)** Original surface, generated from the equation $z = \sin(x) + \sin(y)$ and rescaled to vary between 6 and 14. **(b)** Degraded surface, where 30% of points are randomly chosen values between 5 and 15, and the remaining 70% are corrupted by additive Gaussian noise with $\sigma_e = 0.25$. **(c)** Reconstructed surface given by local maximum likelihood planar patches. All surfaces are displayed as 1/8 resolution mesh plots.

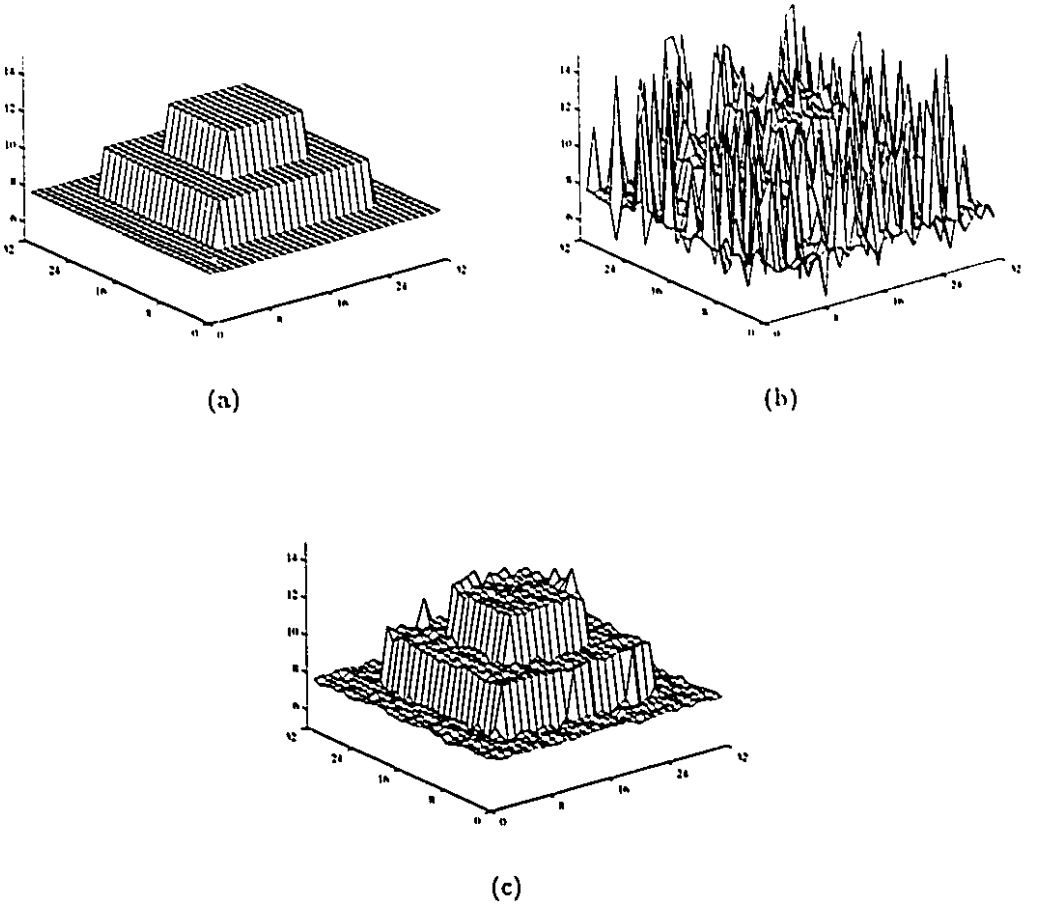


Figure 5.3: Reconstruction of a surface containing discontinuities. **(a)** Original surface, a "wedding cake" arrangement of fronto-parallel planes of height 7.5, 10, 12.5. **(b)** Degraded surface, where 30% of points are randomly chosen values between 5 and 15, and the remaining 70% are corrupted by additive Gaussian noise with $\sigma_e = 0.25$. **(c)** Reconstructed surface given by local maximum likelihood planar patches. All surfaces are displayed as 1/8 resolution mesh plots.

reconstruction technique.

In binocular stereopsis it is common to use a class of artificially created image pairs, called random dot stereograms [42], to test a stereo matching algorithm. An image is created where each pixel intensity is drawn independently from a uniform distribution. A second image is created by shifting the first image by different amounts in different regions, simulating different disparities, and completing the thus unfilled areas with additional random intensities. To use such a stimulus in monocular stereopsis, these two images are simply added together.

In this example, a 256×256 random dot stereogram was created consisting of a central square with disparity 6 standing out from a background of disparity 3 (Fig. 5.4a). The composite image was analyzed by the centering 32×1 windows on each pixel. When such a window overlapped the boundary of the composite image, the disparity value was set to zero. All other windows were analyzed by the cepstrum, with zero-padding to 512 points to reduce aliasing. The resulting raw disparity measurements contain scattered errors throughout, and a fair degree of "jaggedness" around the depth discontinuity (Fig. 5.4b). The scene was reconstructed with 8×8 maximum likelihood planar patches. In the resulting representation (see Fig. 5.4c), not only are the scattered errors in the raw disparity map no longer present, but the discontinuity is localized to within one planar patch or better (Fig. 5.4c). Furthermore, the log likelihood values for each planar patch (represented as grey levels in Fig. 5.4d) clearly indicate the presence and location of the depth discontinuity. These likelihoods can be interpreted as confidence values for each planar facet, and can be input to an even higher level process to interpret the reconstructed surface. It is worthwhile noting that this likelihood, just like the confidence value associated with monocular disparity estimates, is a quantitative probability that can be used in Bayesian or other forms of probabilistic analysis. The input to the surface reconstruction process is a raw disparity map, and associated confidence values; the output is a higher level representation of surfaces in the scene, and associated confidence values.

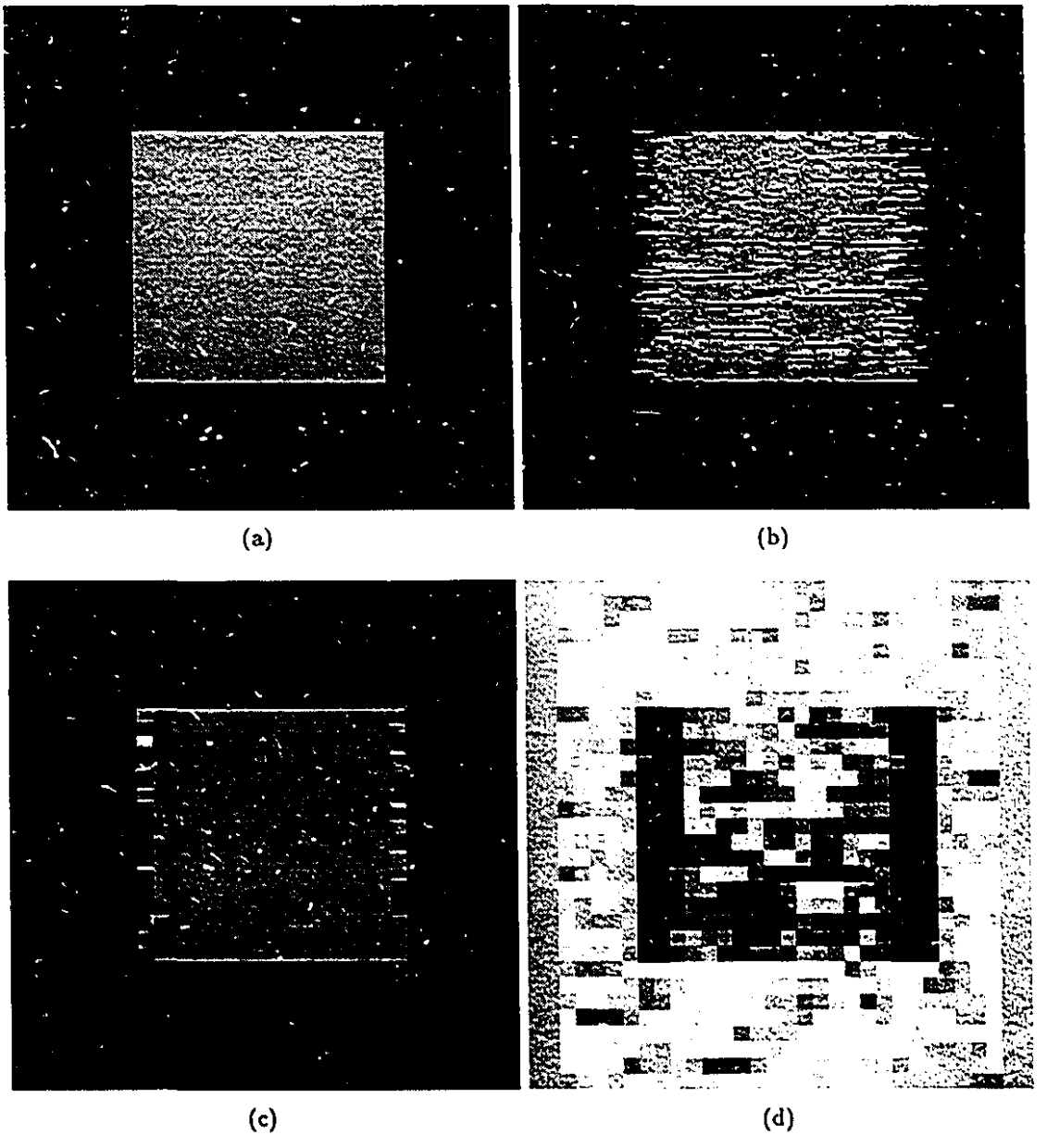


Figure 5.4: Monocular disparity measurement and surface reconstruction for a random dot stereogram. **(a)** Actual disparity values at each pixel of a 256 × 256 composite image, displayed as a grey level image. **(b)** Raw estimated disparity values as determined by application of a 32 × 1 cepstral window to each pixel, except where such a window overlaps the image boundary. **(c)** Reconstructed disparity surface, based on maximum likelihood 8 × 8 planar patches. **(d)** Log likelihood values associated with each maximum likelihood planar patch, rescaled to be displayed as a grey level image.

5.3 Evaluating Spatial Resolution

The technique for converting a composite image into a representation of surfaces in the scene is now complete. Attention can now be turned to evaluating its performance. In computer vision, the most common method for evaluating performance of an algorithm is to present results for various inputs with various parameters. Although Chapter 6 does exactly that, in this section a different approach is taken to evaluating an artificial vision system, by applying the techniques of human visual psychophysics.

The motivation for this approach is twofold. First, psychophysics provides a well developed framework for quantitatively evaluating the capabilities of a vision system. The fundamental tasks of detection, discrimination, and localization are generic to any form of sensory perception. Detection refers to the ability to sense the presence versus the absence of some stimulus, without necessarily being able to identify it. In computer vision, some work has been done to apply techniques from psychophysics to evaluate line detection algorithms [43]. Discrimination refers to the ability to differentiate between two distinct stimuli with different characteristics. Localization refers to the accuracy with which one can judge the position of some stimulus. Second, psychophysics measures performance of a vision system in terms of the subject's *behaviour*. It is therefore closely linked with statistical decision theory [27]. Any autonomous agent must make decisions about its environment based on some form of sensory perception. If the goal is to build a vision system for a mobile robot, what better way to evaluate its performance than to examine the quality of decisions the robot makes about its environment?

Having embraced the framework of visual psychophysics, there are numerous experiments that could be performed. To some extent, the issues of resolution and accuracy of individual depth measurements has been addressed in Sec. 4.4.2. However, the *spatial* resolution (in the x and y coordinate, not the z coordinate) of the range sensor developed in this thesis, has not been evaluated. Depending on the application, the spatial resolution of a range image can be very important in determining

its usefulness. Spatial resolution limits the size of an object that can be detected, and the accuracy with which an object can be localized. Here size and localization refer not to the depth (z) coordinate, but to the x and y spatial coordinates of the range image.

Rather than apply the range sensing technique to various real-world scenes, spatial resolution was measured using carefully designed artificial stimuli. This was to ensure that the fundamental performance of the vision system was being evaluated, not just how it reacted to a specific scene. This is part of the philosophy of visual psychophysics. In fact, any vision system, biological or artificial, can be evaluated by its performance in these tasks.

5.3.1 Obstacle Detection and Discrimination

Suppose a mobile robot is navigating through an unknown environment. Its first priority is to determine if it is safe to continue along its current path. To do so the robot must determine, with a simple yes or no answer, if there is an obstacle directly in front of it. The consequences of a false negative response (e.g., a head-on collision), are more serious than a false positive response (e.g., avoiding an obstacle that is not really there), so a conservative strategy is to always answer yes in the presence of significant uncertainty. The mobile robot requires both the ability to *detect* changes in depth, and to *discriminate* "near" objects from "far" objects. The ability to perform these two tasks as the size of the object is reduced, is a performance characteristic related to the spatial resolution of the range sensor.

To examine this performance an experiment was designed in which an *ideal observer* makes *two-alternative forced-choice* (2AFC) decisions about the presence or absence of a change in depth (detection), and whether this change is a positive or negative one (discrimination). An ideal observer is a procedure that uses the available sensory data (disparity measurements and confidence values) in a statistically optimal manner in order to make a decision. In this case, the ideal observer must choose one of two alternatives, only one of which is correct (hence the 2AFC label). Performance is measured in terms of the percentage of correct responses in a large

number of statistically independent trials. The rationale for the ideal observer is that since the decision is made in an optimal manner, its performance is due to the quality (in this case, spatial resolution) of the sensory data, not the decision procedure.

Because the visual echo is in the horizontal direction only (for horizontally aligned apertures), spatial resolution is higher in the vertical direction than in the horizontal direction. Therefore the focus here is on horizontal resolution. Vertical resolution is determined solely by the vertical window size, as described in Sec. 5.1.2. The use of a 2-D stimulus will also confuse the issue of horizontal resolution. Since information can be integrated over successive scanlines, one would obtain better performance the larger the vertical window size and the larger the vertical extent of the obstacle. To avoid these complications, the experiment will be performed on a 1-D sequence of composite image data.

The stimulus was constructed as follows. A 1-D composite image sequence was formed of length $4N$ (where N is the window length for cepstral analysis) with some monocular disparity d_0. The single image consisted of white noise, the "optimal" texture for visual echo analysis. Within this sequence there are two "fields" of interest, one from position $N/2$ to $3N/2$ (referred to as field A), and the other from position $5N/2$ to $7N/2$ (referred to as field B). The fields are separated so that no N-point window centered in one field will overlap the other field, or the image boundary. One of the two fields contains a sequence (referred to as the obstacle) of length L and disparity d_- or d_+, where $d_- < d_0 < d_+$. The position of the obstacle within the field, and the choice between d_- or d_+, are uniformly randomly distributed. The entire sequence is analyzed by an $N \times 1$ cepstral window centered at every point.

The first task is to decide whether the obstacle occurs in field A or B, given the raw disparities and confidence values at each point. The ideal observer performs this task as follows. One N-point field consists of a flat surface of disparity d_0, the other consists of "something else". The likelihood of the N-point disparity model $D(x) = d_0$ is given by Eqn. (5.3a), with the disparity probability density function given by Eqn. (5.5). Whichever of the two fields had a lower likelihood of this disparity model was taken to be the field containing the obstacle. The percentage of correct responses in 500

statistically independent trials was recorded. This percentage characterizes the ability to *detect* an obstacle of width L/N relative to the window size.

The second task is, given the answer to the first question, to decide whether the obstacle is closer or farther than the background. Suppose, without loss of generality, that the obstacle occurs in field A. Without knowing the width or location of the obstacle within field A, the ideal observer cannot model the actual disparity structure. The field consists of three segments: two with disparity d_0, the other with disparity d_- or d_+. The ideal observer must choose between d_- and d_+. Consider modelling the N-point field as a flat surface of disparity d_- or d_+. The segments at disparity d_0 will contribute equally little to the likelihood of both models, while the other segment will contribute more to the model which matches its disparity. Therefore if a flat surface of disparity d_- has greater likelihood than d_+, the ideal observer concludes the obstacle is closer than the background, and vice versa. After 500 statistically independent trials, the total percentage correct characterizes the ability to *discriminate* an obstacle of width L/N relative to the window size.

This ideal observer assumes prior knowledge of d_- and d_+ in order to make its decision. In practice, under most circumstances these disparity values are *a priori* unknown. It is difficult to formulate an ideal observer in this case, so a heuristic-based "practical observer" was simulated. This observer reconstructs the 1-D disparity profile using 8 point maximum likelihood segments. Since the field contains two depth discontinuities, at least two of these segments will be unreliable. Nonetheless, there should be some segments correctly indicating disparity d_- or d_+, and the mean reconstructed disparity over the N-point field should be biased relative to d_0 accordingly. Therefore if this mean is less than d_0, the practical observer concludes the obstacle is closer than the background, and vice versa. Note that this observer is not ideal because the distribution of errors in reconstructed disparity values cannot be assumed to have zero mean.

The experiments described above were repeated for different values of obstacle length L, ranging from $L = N$ to $L = N/16$, with a window length of $N = 128$ (see Fig. 5.5). These results are for ideal stimuli — an unblurred, noise free, random dot

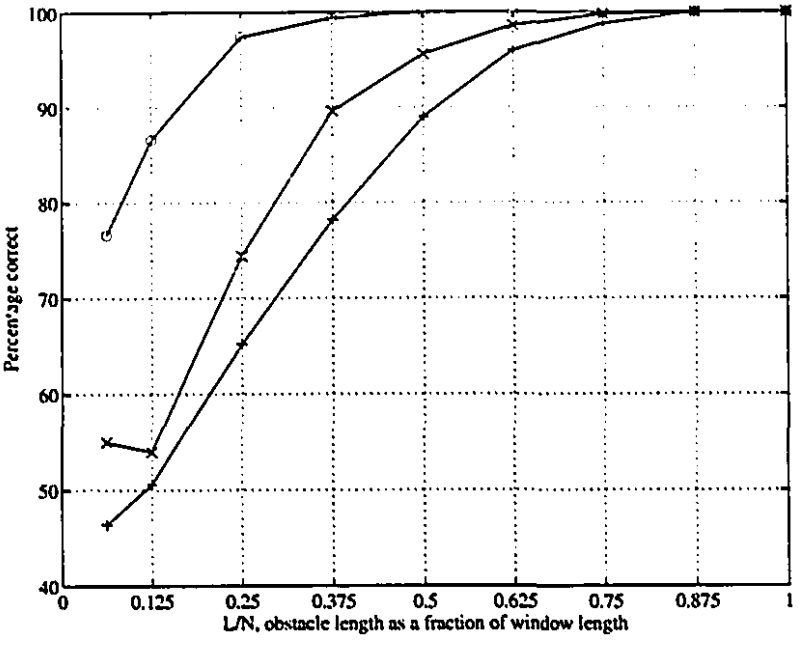


Figure 5.5: Obstacle detection and discrimination for ideal stimuli. In the detection task (the top curve with points indicated by o), the ideal observer must choose which of two fields contains an obstacle of width L. In the discrimination task, the observer must decide whether the obstacle is closer or farther than the background. Two discriminators are shown, one the ideal observer (the x points), the other more typical of how the range sensor would behave in practice (the + points). The stimuli are unblurred, noise-free random dot patterns.

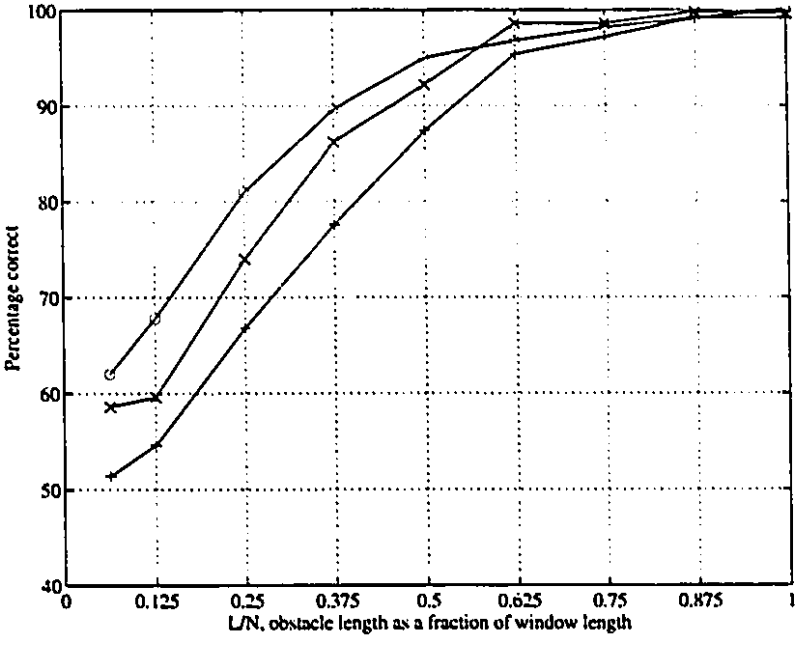


Figure 5.6: Obstacle detection and discrimination for degraded stimuli. The same three curves as in Fig. 5.5, except in this case each stimulus was blurred with a 1–D Gaussian kernel of width $\sigma_b = 1$, and Gaussian noise added at a SNR of 40dB.

	Obstacle width threshold (as a fraction of window length)	
	Ideal stimuli	Degraded stimuli
Ideal Detector	0.06	0.19
Ideal Discriminator	0.25	0.28
Practical Discriminator	0.34	0.36

Table 5.1: Obstacle width thresholds for detection and discrimination. These thresholds are given by the obstacle size required to obtain 75% correct responses in the experiments depicted in Figs. 5.5 and 5.6.

pattern. To simulate performance under more typical conditions, the experiments were repeated with each stimulus blurred by a 1-D Gaussian kernel of width $\sigma_b = 1$, and Gaussian noise added at a SNR of 40dB (see Fig. 5.6). This represents a significant amount of noise and blur (in the horizontal direction), more than would be expected from a good quality camera with narrow, vertical slit apertures.

To obtain quantitative performance limits from these plots, a threshold performance level of 75% correct is chosen. The obstacle width (as a fraction of window size) corresponding to a level of 75% correct is referred to as the obstacle width threshold for the particular task (see Table 5.1). The lower the threshold, the higher the effective spatial resolution of the range sensor. In the detection task, there is a significant difference in performance between ideal and degraded stimuli. Blur and noise tends to obscure narrow obstacles, while introducing significant errors in the obstacle-free field of each stimulus. Nonetheless, given the results of these experiments, it can be said that under most circumstances, obstacles as narrow as one-eighth the window length can be reliably detected. The discrimination task is generally more difficult than the detection task, requiring a larger obstacle width in order to be successful. This is because high confidence, "correct" disparity estimates are required to reliably discriminate between near and far depths. It can be said that under most circumstances, obstacles as narrow as one-third the window length can be reliably discriminated.

5.3.2 Depth Discontinuity Localization

Once an obstacle has been detected and identified as being close to the robot, the next task is to determine the position of the obstacle so as to manoeuvre around it. This task is referred to as localization. If the robot decides that an obstacle is located between positions x_1 and x_2, it should avoid the region between $x_1 - \epsilon$ and $x_2 + \epsilon$, so as to allow for uncertainty in spatial measurements. The choice of ϵ is determined by the spatial resolution of the range image. The goal in this experiment is to estimate the value of ϵ as a fraction of the cepstral window length.

To formulate the localization task as a 2AFC experiment, the ideal observer will be asked to determine if an obstacle is to the right or left of the centre of the composite image. Since an obstacle is defined by a discontinuity in depth, the stimulus will consist of a step change in disparity, rather than an object of different disparity than its background. Whenever there is a step change in disparity, there is an interruption in the visual echo — a short sequence of points that have no echo (see Fig. 5.8). A small region in the scene immediately adjacent to the discontinuity is visible from only one of the two apertures, referred to as a partially occluded region. The greater the difference in disparity across a discontinuity, the larger the partially occluded region. Because of this there is some uncertainty as to what constitutes the "true" location of a disparity discontinuity in the composite image. This must be addressed in order to evaluate outcomes of the 2AFC experiment. First the stimulus is described, and the nature of the ideal observer.

As in the obstacle detection and discrimination experiment, the stimulus consisted of a 1–D composite image sequence, where the single image consisted of white noise. As before, the sequence had length $4N$, where N is the window length. The position x_d of the discontinuity was chosen at random from N to $3N$, defined as the point at which disparity changes from d_- to d_+, where $d_+ > d_-$. The entire sequence was analyzed by an $N \times 1$ cepstral window centered at every point. The task for the ideal observer was to first form an estimate, $\hat{x_d}$, of the position of the discontinuity. If $\hat{x_d} < 2N$ the discontinuity was labelled L for left of centre, otherwise it was labelled R for right of centre. After a large number of statistically independent trials, the

percentage of estimates labelled R was plotted as a function of the true discontinuity position.

The ideal observer for this task must determine the maximum likelihood position of the step edge. Therefore instead of using local planar facets, the appropriate disparity model (for the entire stimulus) is given by

$$D(x;\, x_d) = \begin{cases} d_- & \text{if } x < x_d \\ d_+ & \text{if } x \geq x_d \end{cases} \tag{5.7}$$

where d_- and d_+ are the monocular disparity values on either side of the discontinuity. The global maximum of the likelihood function for this disparity model gives the maximum likelihood position of the step edge. For a more general model, the disparity values d_- and d_+ can be determined as model parameters, but in such a case the likelihood function will be very complex in terms of multiple local maxima.

As an illustration of the step localization process, the results of analyzing three different step edge stimuli are given in Fig. 5.7. The first column displays a profile of raw disparity values as given by a 128-point cepstral window centered on each point. The true disparities are 10 and 15, while the disparity range tested was 5 to 20. Notice that in some cases the raw disparities do resemble a step edge, while in other cases there are numerous incorrect measurements (that are neither 10 or 15) in the vicinity of the discontinuity. The second column shows the probability correct values corresponding to the disparity profile in the first column. As expected, when the cepstral window significantly overlaps a discontinuity, the resulting confidence value is much lower than in the constant disparity case. The third column shows the log likelihood function for the disparity model given in Eqn. (5.7) (with a negative constant term removed, therefore it may exceed one). The location of the maximum of this function, indicated by the dashed vertical line, is the ML location of the step edge as determined by the ideal observer.

Before performing the 2AFC experiment, the "true position" of a step edge in disparity must be defined. For example, Fig. 5.8 depicts the formation of a composite image sequence in which monocular disparity jumps from $d_- = 3$ to $d_+ = 6$. The

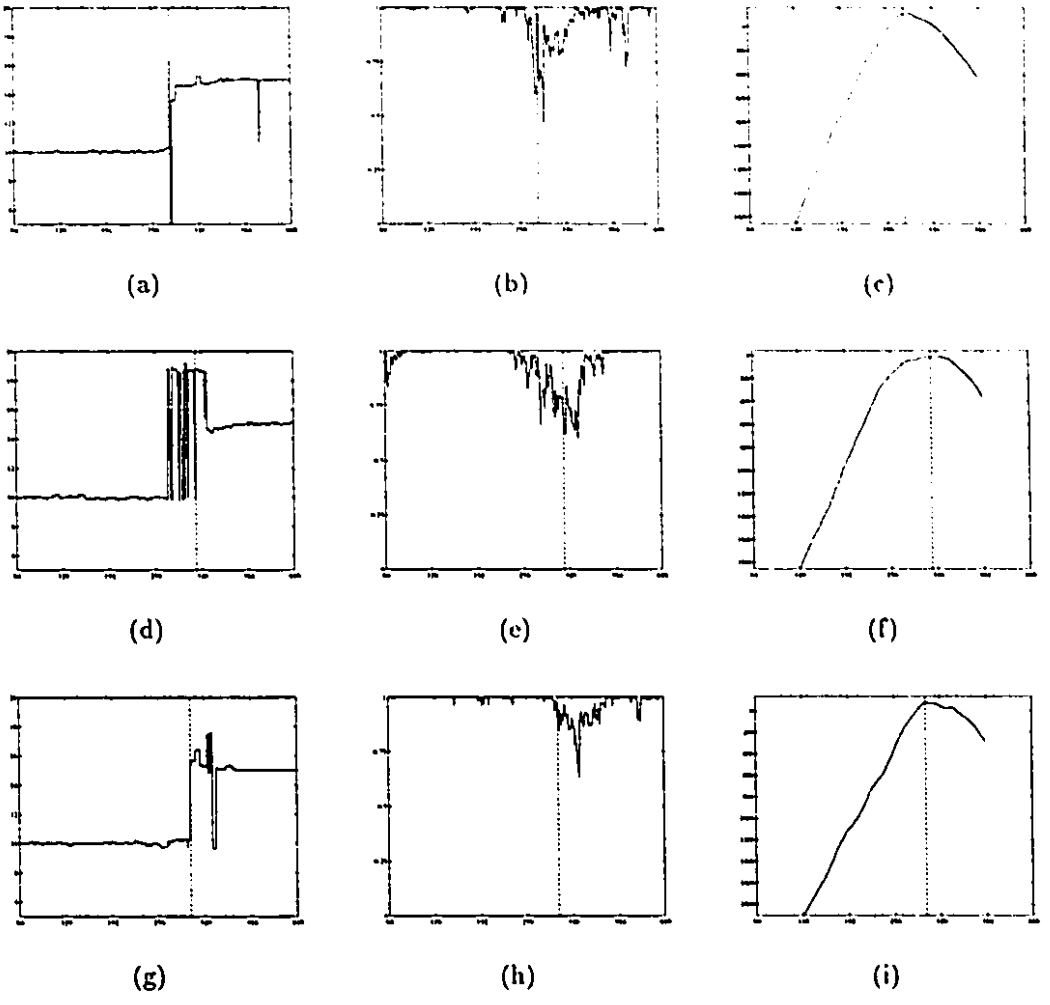


Figure 5.7: Maximum likelihood localization of a step edge in disparity. **(a,d,g)** Raw disparity measurements using a 128-point cepstral window. **(b,e,h)** Estimated probability that the correct cepstral peak was selected, corresponding to each measurement in column one. **(c,f,i)** Log likelihood function (plus a constant) for a step edge disparity model, the parameter of which is the location of the step edge. The maximum value of this function gives the maximum likelihood location of the discontinuity (indicated by a dashed vertical line in each column).

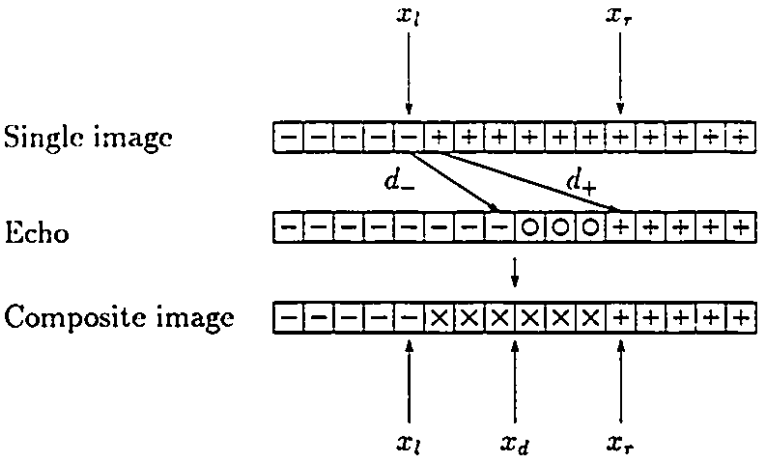


Figure 5.8: Formation of the composite image around a depth discontinuity. In the composite image, the discontinuity "begins" at position x_l and "ends" at position x_r. The true position of the discontinuity is given by $x_d = (x_l + x_r)/2$. The "–" symbols indicate points with disparity $d_- = 3$, "+" indicates points with disparity $d_+ = 6$, and "o" indicates partially occluded regions.

right-most position in the composite signal with disparity d_- is denoted by x_l. The left-most position in the composite signal with disparity d_+ is denoted by x_r. Between x_l and x_r, disparity in the composite signal is not well defined. Any of these locations may be chosen by the ML observer as the true location of the discontinuity. The median of these chosen positions is expected to be the centre of the region between x_l and x_r, $(x_l + x_r)/2$.

An experiment was performed to confirm this hypothesis. The ideal observer described above was used to predict the location of a discontinuity, held in a fixed position over a large number of trials. In particular, referring to the symbols used above, in each stimuli $d_- = 5$, $d_+ = 15$, $x_l = 192$, $x_r = 208$. A histogram of ML step locations in 4,000 trials (with a window size of 128 points) was computed (see Fig. 5.9). A Gaussian distribution with mean and variance given by the observed distribution is superimposed on the data. The mean and median of this distribution are nearly equal, at 201, roughly $(x_l + x_r)/2$ as expected. Also notice that over the

interval $[x_l, x_r]$, the distribution of ML step locations is more uniform than Gaussian. Nonetheless, this provides a definition for the true position of a step edge, given by $(x_l + x_r)/2$.

The horizontal axis in Fig. 5.9 is in units of absolute horizontal position. However, as in the detection and discrimination experiments, localization ability is strongly dependent on window size. Therefore localization error is better expressed as a fraction of window length. In Fig. 5.9, the standard deviation of localization error expressed in these units is 0.17. However, in this experiment a large disparity step was used in order to analyze the true position of the step edge. The larger this disparity step, the larger the partially occluded region and the more error introduced in the localization task.

In the 2AFC experiment, a smaller disparity step from 8 to 12 is used. The results of this experiment consist of a L/R response and a true position (as given by the above definition) of the discontinuity, for each of 4,000 statistically independent trials. To present these results as a percentage correct, this data was sorted by true position and collected into bins. These true position values are labelled so that zero corresponds to the centre of the composite image. Within each bin, the percentage labelled R is plotted against the true position of the discontinuity. The results are given in Fig. 5.10. As expected, if the discontinuity occurs on the far left, there are 0% R responses, while if the discontinuity occurs on the far right, there are 100% R responses. It is the transition between these two extremes that characterizes localization performance. Superimposed on the data is the best fitting smooth "psychometric function", a function of the form in Eqn. (4.18). Signal detection theory predicts that the outcome of such an experiment will follow such a curve [27].

To simulate performance under more adverse conditions, the entire experiment was repeated with each stimulus blurred by a 1-D Gaussian kernel of width $\sigma_b = 1$, and Gaussian noise added at a SNR of 40dB (see Fig. 5.11). For both the ideal and degraded stimuli, adopting a 75% correct performance criterion, discontinuity localization is possible to within 1/8 of the window length. Notice that the discontinuity localization task is much less sensitive to noise and blur than the detection

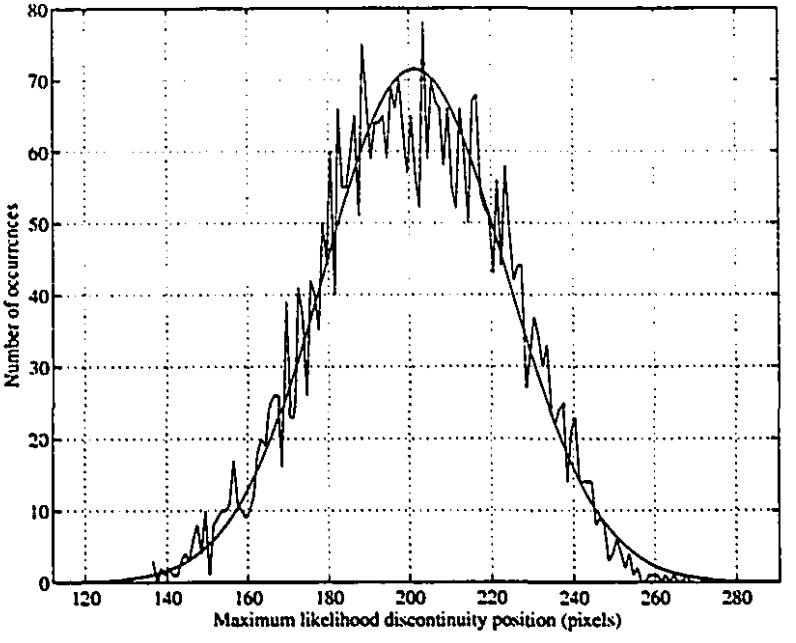


Figure 5.9: Histogram of maximum likelihood step locations. A step edge with $x_l = 192$, $x_r = 208$ is placed in a 512-point composite image sequence (consisting of white noise) and analyzed by a 128-point cepstral window. The maximum likelihood step location is recorded in 4,000 independent trials to form the above histogram. A Gaussian calculated from the parameters of this distribution (mean 201.28, standard deviation 22.31) is superimposed on the histogram.

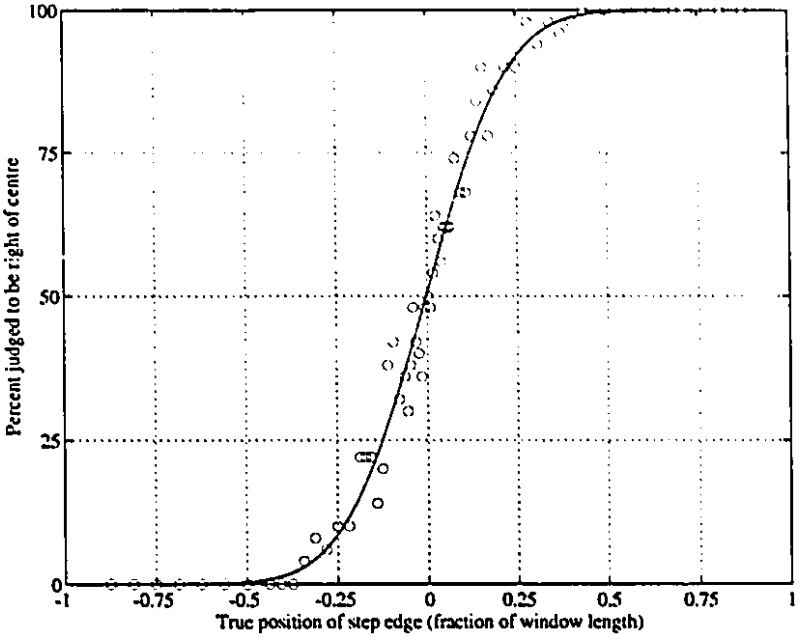


Figure 5.10: Localization of a step edge in disparity for ideal stimuli. The outcome of a 2AFC experiment in which the ideal observer must determine if a step edge in disparity is to the left or right of a centre point. The horizontal axis indicates the true position of the discontinuity, normalized by window length and shifted so that zero corresponds to the centre. The points indicate the percentage of trails in each bin that were judged to be right of centre. The best fitting psychometric function is superimposed on these data points.

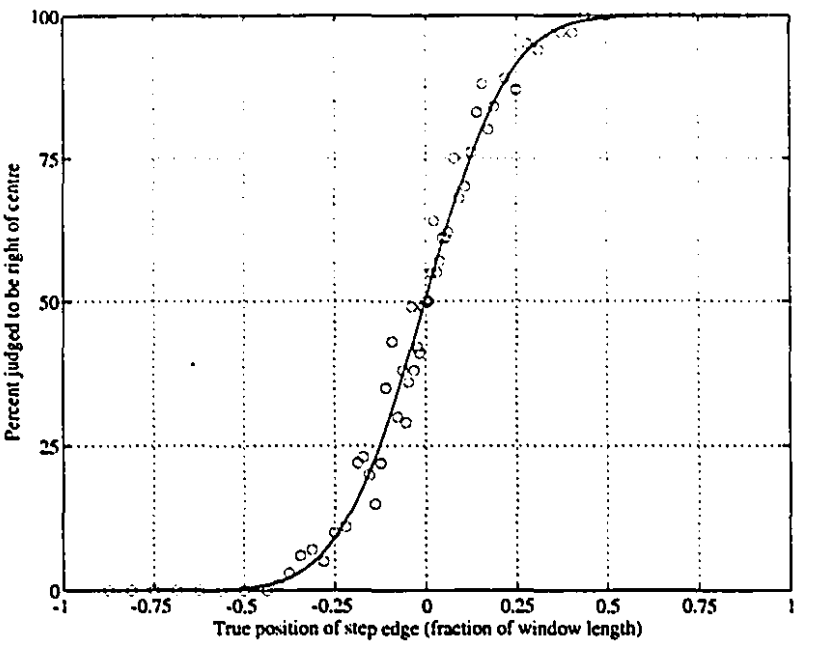


Figure 5.11: Localization of a step edge in disparity for degraded stimuli. The same curve as in Fig. 5.10, except in this case each stimulus was blurred with a 1–D Gaussian kernel of width $\sigma_b = 1$, and Gaussian noise added at a SNR of 40dB.

and discrimination tasks. Blurring has the effect of smoothing the sharp transition in disparity at the discontinuity, but does not significantly obscure the location of the discontinuity.

5.4 Summary

In the composite image acquired by a multiple aperture camera, the 3-D structure of a scene is encoded by monocular disparity. Cepstral analysis allows the detection of this disparity, at each point over an image sampling grid. Measurements of depth (in the form of monocular disparity) over windows of the composite image are recorded at the centre of each window, along with estimates of the error distribution associated with each measurement. However, in this result the 3-D structure is not yet made explicit. The surface reconstruction framework proposed in this chapter enables the conversion of this data into a higher level, model-based description of the 3-D world as seen by the camera. Using this framework, the fundamental visual tasks of detection, discrimination and localization can be studied in terms of the decisions of an ideal observer. The thresholds on obstacle width for detection and discrimination are 1/8 and 1/3 the window length, respectively, while depth discontinuity localization is accurate to within 1/8 the window size.

Chapter 6

Experimental Results

The range sensing technique developed in this thesis was applied to composite images of real-world scenes. These scenes were chosen to reflect different applications of range imaging, where different tasks were to be performed by an intelligent machine on the basis of a range image. The sensor was qualitatively evaluated in terms of the ability of the machine to complete the required task.

This approach allows a different kind of performance evaluation, compared to just presenting a series of arbitrarily chosen range images and asking the reader to judge their apparent quality. Given the resolution and accuracy with which we humans view the world, a pictorial representation of how a machine views its environment may seem very crude by comparison. Instead, we should take a step back and ask ourselves what task is to be performed by the machine, and what it *needs* to know about its environment in order to achieve this task. Any sensory information that is over and above what is required to complete the desired task, is superfluous, representing unnecessary computational expense.

6.1 Procedure Used to Acquire and Process Images

The composite images presented in this chapter were acquired as follows. A double aperture mask was constructed from a very thin brass disc (although any opaque

material would have been sufficient). Two openings of identical size and shape were created, equally spaced about the centre, along a diameter of the disc. Two such masks were constructed, one with 0.5 mm diameter pinholes, the other with 0.5 mm × 4.0 mm slits. In each case, the distance between the centre of the two apertures was 6.0 mm (i.e., in Fig. 3.1, $D = 6.0$ mm, $A = 0.5$ mm).

Two different cameras were used in these experiments. The first was a black and white CCD (charge coupled device) camera with a standard 16 mm television lens. In this camera, the mask was mounted between the lenses, directly behind the fully open iris diaphragm. The second camera was a SLR (single lens reflex) 35 mm film camera with a standard 55 mm lens. Here the mask was placed directly behind the lens, between the lens and the shutter. In general, the mask should be positioned as close as possible to the iris (i.e., at or near the effective lens centre), centered on the optical axis of the camera. The greater the focal length of the lens, the less critical the mask position relative to the iris. However, the mask must be positioned in such a way that when the focus knob is adjusted, the mask moves along with the lens, relative to the image plane. This allows the range of monocular disparities present in a composite image to be controlled by adjusting the depth at which the camera is focused.

Before acquiring an image with either camera, it was focused either in front of or behind the 3-D objects of interest. This was to ensure that all points in the scene had a non-zero monocular disparity value, and that all disparities were of the same sign. The depth at which the camera was focused (readable directly from the lens body), and whether objects of interest were closer or farther than this depth, was recorded. This information is necessary to convert measured disparities (unsigned quantities) to real 3-D distance. To achieve sufficient luminance and contrast in the composite image, a high level of scene illumination or long exposure time was used, but not to the extreme that the composite image became saturated.

Pictures from the CCD camera were digitized directly to 640 × 480 8-bit grey level images, on a computer workstation. These images were immediately ready for processing by the cepstrum. For the SLR camera, the 35 mm color film was developed

Composite Image	Z (m)	D (m)	Aperture shape	Image Size horiz.	Image Size vert.	Disparty Range min	Disparty Range max	Disparty Range sign
Concrete steps	0.5	.0082	· ·	1536	1024	25	39	−
Tulip bed (R/G/B interlaced)	0.5	.0085	· ·	1536	3072	20	45	−
Tree and sculpture	1.0	.0081	· ·	1536	1024	9	22	−
Toy Godzilla	0.35	.0098	| |	1536	1024	9	45	−
Tabletop, 2 objects	0.38	.010	| |	1536	1024	7	22	−
Tabletop, 4 objects	0.77	.006	· ·	640	480	7	15	+
Robot view 1	∞	.0075	| |	1536	1024	8	30	+
Robot view 2	∞	.0074	| |	1536	1024	10	32	+
Robot view 3	0.7	.0080	| |	1536	1024	12	18	−
Robot view 4	1.0	.0088	| |	1536	1024	18	30	+
Robot view 5	0.9	.0080	| |	1536	1024	12	20	−

Table 6.1: Parameters for composite image acquisition. From left to right, the parameters listed are: the depth at which the camera was focused, the effective aperture separation, the aperture shape (pinholes or slits), the resolution of the composite image, and the range and sign of monocular disparities in the composite image. The composite images themselves are given in Figs. 6.1–6.13.

and digitized to 3072 × 2048 24-bit colour images. These images were converted to 1536×1024 or 768×512 8-bit grey level images for processing. The optical parameters, image sizes, and disparity ranges for the composite images analyzed in this chapter are given in Table 6.1.

One of the advantages of the cepstral technique of monocular disparity measurement developed in this thesis, is the absence of arbitrarily chosen thresholds or parameters that have a significant impact on performance. For a given composite image, once the approximate range of disparities has been estimated (either by visual inspection, or from prior knowledge of the focus setting and range of depths to be encountered), the remaining parameters are chosen to trade-off speed for resolution and accuracy. These parameters are: horizontal and vertical window size, extent of zero-padding, and disparity map density. For example, assume the range of disparities in the scene is $d_{min}, ..., d_{max}$. For maximum speed, choose a window size of $4d_{max} \times 1$, zero padded up to 2^k, where $k = \lceil \log_2(4d_{max}) \rceil$ (the total length of the

sequence after zero-padding must be a power of two for the FHT algorithm), and a low disparity map density, such as $(1/d_{max}, 1/4)$. The resulting disparity map will have low resolution in the horizontal direction, and contain some errors due to the small window size. A higher density disparity map provides greater resolution, and a larger window size and more zero-padding will reduce disparity errors. Depending on the application, the improved result may be worth the price in additional computation. The particular parameters used in the experiments described in this chapter are given in Table 6.2.

For the surface reconstruction technique described in this thesis, the input parameters consist of the choice of a local surface model, and the dimensions of the composite image region to be approximated by one instance of this model. In particular, if a piecewise planar facet model is chosen, the dimensions in pixels of each rectangular planar facet are the only parameters that must be specified. The larger the facet, the more data is available on which to base the fit, but the more likely it is that the true surface deviates from a planar model over the area of the facet. The particular planar facet dimensions used in the experiments described in this chapter are given in Table 6.2.

The final step is the conversion of the reconstructed surface from disparity space to 3-D space (i.e., converting disparity to depth). Knowing the focal length, F, and the depth at which the camera is focused, Z, the lens to sensor plane distance, f, is recoverable from the Gaussian lens equation given in Eqn. (3.1). The distance between the two apertures, D, is measurable directly from the mask inserted into the camera. However, the geometric optics formulation in Sec. 3.1 assumes the mask is placed in the effective centre of the camera lens. In the cameras described above, it was placed in front or behind the lens. For a given double aperture camera setup, a calibration procedure is required to measure the *effective* D, the equivalent aperture separation if a mask were to be placed at the lens centre. This effective aperture separation tends to vary with focal setting and other optical properties of the particular camera. The values of effective D determined during the experiments described in this chapter are given in Table 6.1.

Composite Image	Window Size			Map Density		Facet Size	
	horiz.	vert.	total	horiz.	vert.	horiz.	vert.
Concrete steps	180	5	1024	1/2	1/2	32	32
Tulip bed (R/G/B interlaced)	256	15	4096	1/2	1/6	16	16
Tree and sculpture	98	20	2048	1/2	1/2	4	4
Toy Godzilla	200	5	1024	1/2	1/2	4	4
Tabletop, 2 objects	92	16	2048	1/2	1/2	n/a	n/a
Tabletop, 4 objects	80	12	1024	1	1	n/a	n/a
Robot view 1	128	16	2048	1/2	1/2	16	16
Robot view 2	128	16	2048	1/2	1/2	16	16
Robot view 3	128	16	2048	1/2	1/2	16	16
Robot view 4	128	16	2048	1/2	1/2	16	16
Robot view 5	128	16	2048	1/2	1/2	16	16

Table 6.2: Parameters for composite image processing. From left to right, the parameters listed are: the dimensions of the composite image windows extracted for cepstral analysis, the total length (including zero-padding) of the sequence input to the cepstrum, the density of the disparity map computed (expressed as the ratio of disparity map dimensions to composite image dimensions), and the dimensions of the planar facets used to reconstruct the scene. The results of processing these images with these parameters are given in Figs. 6.1–6.13.

The monocular disparity values computed from the cepstrum must be expressed in units consistent with other parameters (e.g., mm). This conversion factor was determined by dividing the horizontal size of the image plane (CCD array or film) by the horizontal image resolution. The disparity values were then given a sign: negative if the camera was focused in front of the scene (i.e., all scene points are at a depth greater than a reference plane, the image of which is in focus), positive if the camera was focused beyond the scene (i.e., all scene points are at a depth less than the reference plane). For each composite image point, $P'(i,j)$, depth can be calculated directly from Eqn. (3.4b), providing the z coordinate of the corresponding point, $P(x,y,z)$, in the scene. If required, the x and y coordinates of P are given by

$$x = u_i \frac{z}{f} \tag{6.1a}$$

$$y = v_j \frac{z}{f} \tag{6.1b}$$

where (u_i, v_j) are the image plane coordinates of the midpoint of the line joining the P' and its echo. This completes the process used in these experiments to acquire a composite image and convert it into a representation of 3-D structure.

6.2 Recovery of Terrain Structure

Consider the task of recovering the basic 3-D structure of the terrain in front of a stationary viewer. This is a task that we humans must perform regularly in order to move freely about our environment. While moving, we avoid collisions with walls or furniture, and we can find our way through doors, around corners, or up and down stairs. As effortless as they may seem, all of these tasks require complex sensory processing.

A composite image taken with a 35 mm SLR camera with two pinhole apertures is shown in Fig. 6.1a. The camera was mounted horizontally on a tripod at the landing of a set of exterior, concrete steps. The two apertures were aligned horizontally, parallel to the eventual scanlines of the composite image. The scene from this viewpoint

consisted of a series of horizontal and vertical planes, the risers and treads of the steps. The darker areas on the extreme left and right of the composite image correspond to regions of the film upon which light is cast from only one of the two apertures. In these areas there is no visual echo cue for depth, therefore they are masked out in the disparity and range images presented throughout this chapter. The camera was focused at a point just in front of the bottom of the steps, with disparity ranging from a minimum of 24 at the bottom to a maximum at 39 at the top of the composite image.

The 1536×1024 composite image was processed by computing the cepstrum of 180×5 sliding windows, with a step size of 2×2, to produce a 768×512 disparity map (i.e., a disparity map density of $(1/2, 1/2)$). Displaying the raw disparity map as a normalized grey level image clearly reveals the basic 3-D structure of the scene (Fig. 6.1b). Based on this disparity map and the associated confidence values, the scene was reconstructed (in disparity space) using 32×32 maximum likelihood planar patches. The resulting surface representation was then converted to 3-D spatial coordinates, and transformed into a global frame of reference for a more intuitive presentation of the structure of the scene (Fig. 6.1c). The few significant errors in the disparity map do not appear in the final result, indicating the effectiveness of the surface reconstruction procedure. The position of the camera from which the composite image was acquired is indicated to the right of the mesh plot. Based on this result, a mobile robot with sufficient dexterity could easily traverse the stairs. Similar results were obtained at lower computational cost, by using a lower resolution image, smaller window size, or lower density disparity map.

The second scene consists of a bed of tulips as seen from an oblique, near horizontal viewing angle (Fig. 6.2a). The structure of this scene can be described in terms of two components. First, there is the receding ground plane characteristic of viewing any horizontal surface at an oblique angle. Second, there are individual tulip flowers protruding up from the green foliage. Excluding the tulips, the depth map should vary linearly, from near at the bottom to far at the top of the image, while containing small isolated patches that are generally nearer than their immediate surroundings.

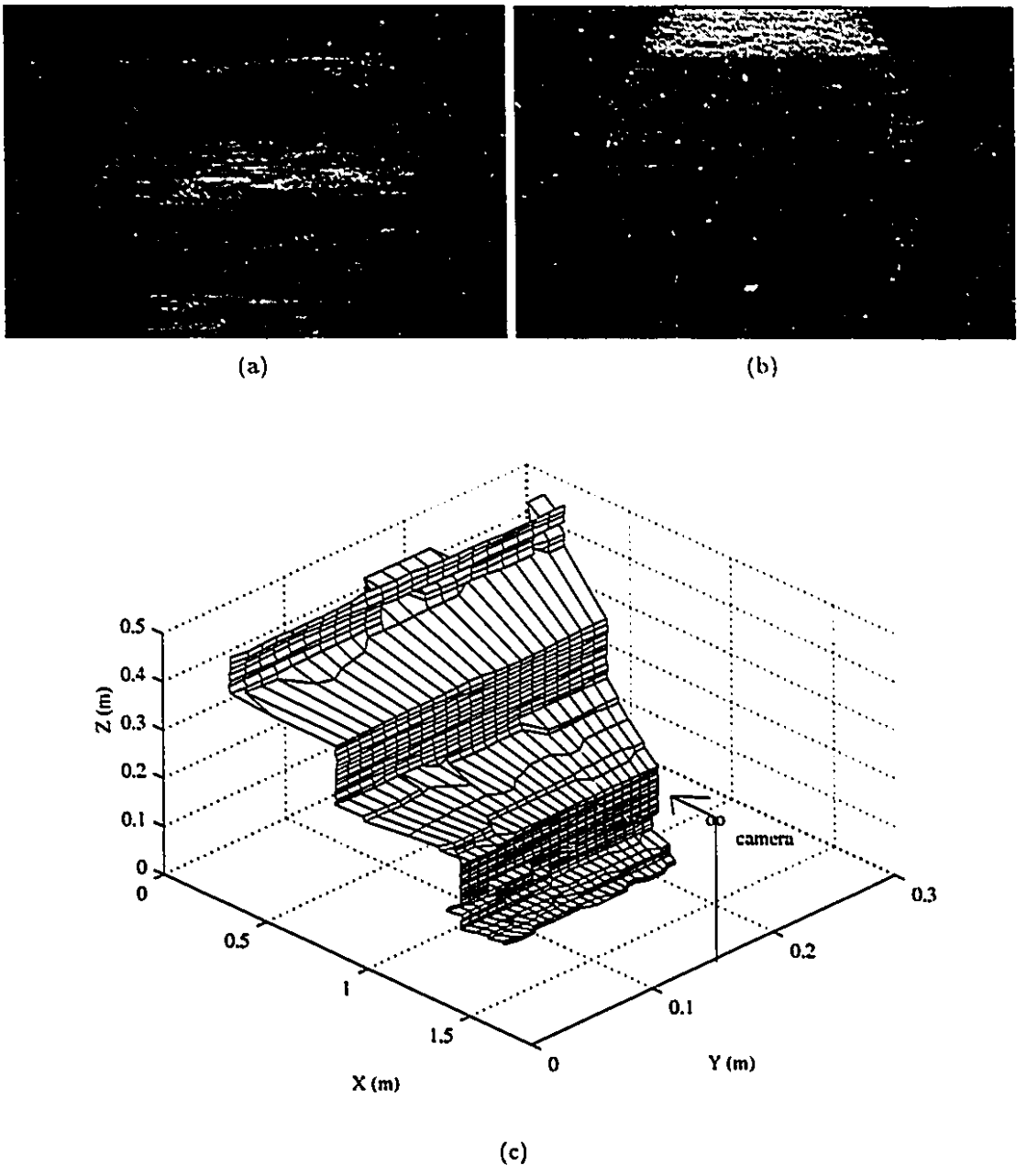


Figure 6.1: Scene of a set of exterior concrete steps. **(a)** A composite image taken from the landing of a set of outdoor steps made of concrete. **(b)** The raw monocular disparity map as provided by centering cepstral windows on one-quarter of the composite image pixels. The disparity value at each pixel is represented as a grey level, where dark intensities correspond to smaller disparities (closer to the camera) while bright intensities correspond to larger disparities (farther from the camera). **(c)** Mesh plot of steps in 3-D global coordinates. Local maximum likelihood planar patches were fit to the raw disparity map, which in turn were converted into planar patches in depth. The resulting surface points in 3-D coordinates are displayed in a global coordinate frame, in which the camera viewpoint is also indicated.

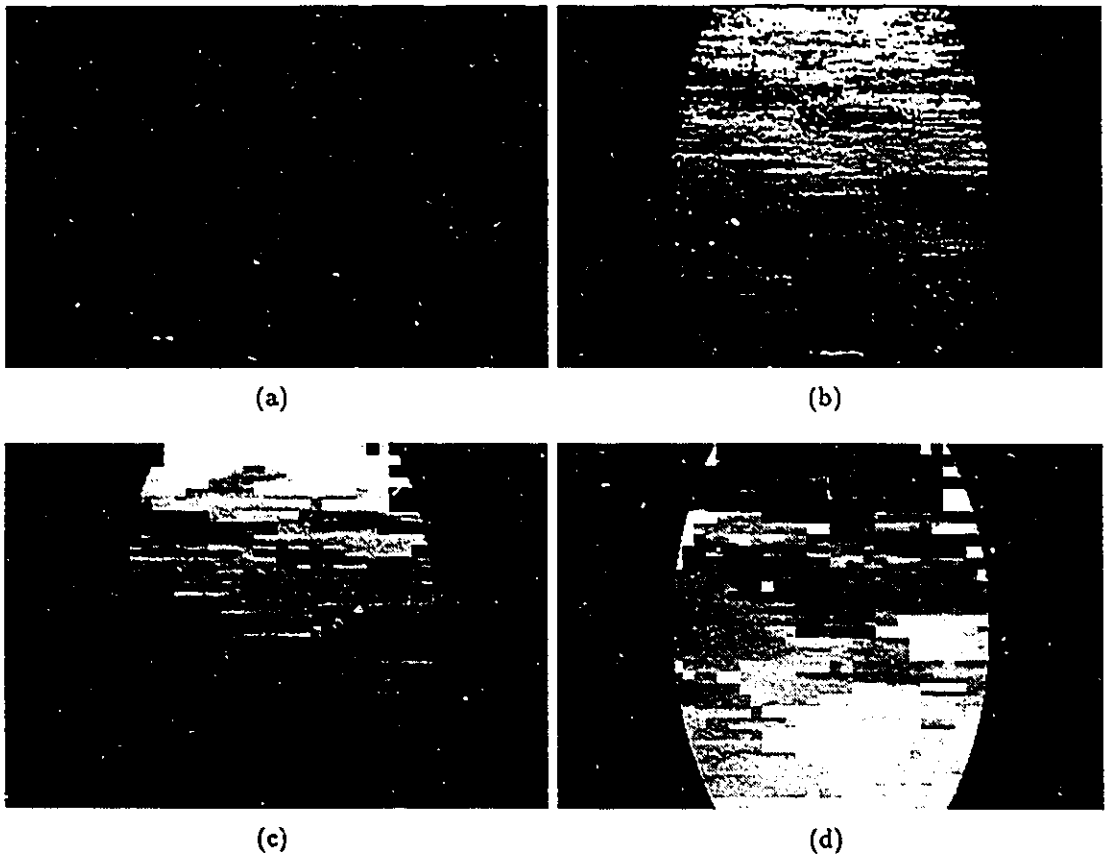


Figure 6.2: Scene of a bed of tulips. **(a)** A composite image of a bed of tulips as seen from an oblique angle. **(b)** The raw monocular disparity map provided by cepstral analysis of the R/G/B scanline interlaced composite image. Dark intensities represent small disparities, while bright intensities represent large disparities. **(c)** Reconstructed surface in disparity space (displayed as a disparity map), given by fitting maximum likelihood local planar patches to the raw disparities. **(d)** Range image provided by converting the disparity map of (c) into real depth. The darker the intensity of a pixel in this image, the farther that point from the camera.

This scene was also viewed with the 35 mm, twin-pinhole camera, but to exploit the rich colour of the tulip bed, the image was prepared in a slightly different manner. Rather than convert directly from a 24-bit colour to 8-bit grey level image, the red (R), green (G), and blue (B) components of the colour image were extracted separately. The colour components were then recombined by interlacing R/G/B scanlines to form an image three times the vertical size of the original. This image was analyzed using a window size three times larger (in the vertical direction) than normal, and a sliding window step size three times larger in the vertical direction. In this way, the disparity map maintained the same horizontal to vertical size ratio as the original image.

The advantage of the R/G/B interlacing technique is that three separate, potentially independent channels of composite image data, with exactly the same visual echo in each, are available to the cepstrum. In other words, "colour of origin" information is not lost. When a colour composite image is converted into a black and white image, points of different colour may be mistakenly interpreted as echoes of one another. If the three colour channels are not collapsed into one, this potential problem can be avoided.

The ability of this technique to improve performance is limited by the extent to which the colour channels are truly independent. In natural images, two or more of the three colour channels tend to covary [16]. One solution to this problem is to use three light sources, each projecting an independent texture pattern onto surfaces in the scene, but in different coloured (red, green and blue) light.

After interlacing the three colour channels, the 1536 × 3072 composite image was processed with 256 × 15 sliding cepstral windows applied with a step size of 2 × 6 pixels. Even with colour interlacing, this scene is much more challenging to process than the steps of the previous example. In many areas, disparity changes rapidly from one horizontal position to the next, that is, there are few areas where disparity is constant over an entire window region. Furthermore, there is very little texture among the leaves and flowers themselves, on which to base an estimate of the visual echo delay. Nonetheless, the raw disparity map (Fig. 6.2b) reveals both the receding ground plane and the tulips in the foreground. After fitting 16 × 16

maximum likelihood planar patches to this disparity map, the 3-D structure of the scene is more apparent (Fig. 6.2c). This local surface representation is converted from disparity into 3-D space, then sampled at image grid locations to provide a range image for display purposes (Fig. 6.2d). In this image, the tulips at the bottom of the composite image are seen to be much closer than those at the top, an expected result due to the near horizontal viewing angle.

6.3 Obstacle Detection

A somewhat more sophisticated task than recovering basic terrain structure, is to detect and localize an obstacle in space in order to avoid a collision. For example, consider the task of running through a dense tree forest. The runner is not concerned with fine surface detail like knots on the tree trunks. The primary concern is to avoid a head-on collision! In terms of visual perception, the required task is to detect any objects that lie directly ahead, and are close enough to necessitate an immediate change in course.

An example of a scene that may arise in such an application is shown in Fig. 6.3a. The tree on the left was quite close to the camera ($\approx$ 1.5 m), while the rectangular sculpture was more distant ($\approx$ 8 m) and the building in the background much farther ($\approx$ 30 m). The SLR camera with twin-pinhole apertures was focused at a depth of 1.0 m (closer to the camera than the tree), so that disparities in the scene are all negative. Due to the nonlinear relationship between monocular disparity and depth (see Fig. 3.2), the sculpture and the building have very similar disparity values (a difference of only 2 pixels at 1536 $\times$ 1024 resolution) despite their large difference in depth. This example further illustrates the importance of making precise measurements of disparity, so that objects at different depths can be discriminated. The raw disparity map (Fig. 6.3b) clearly reveals the tree standing out on the left, and the rectangular sculpture of slightly lower disparity than the background. Some areas of the raw disparity map contain a large number of significant errors, enough to warrant some discussion.

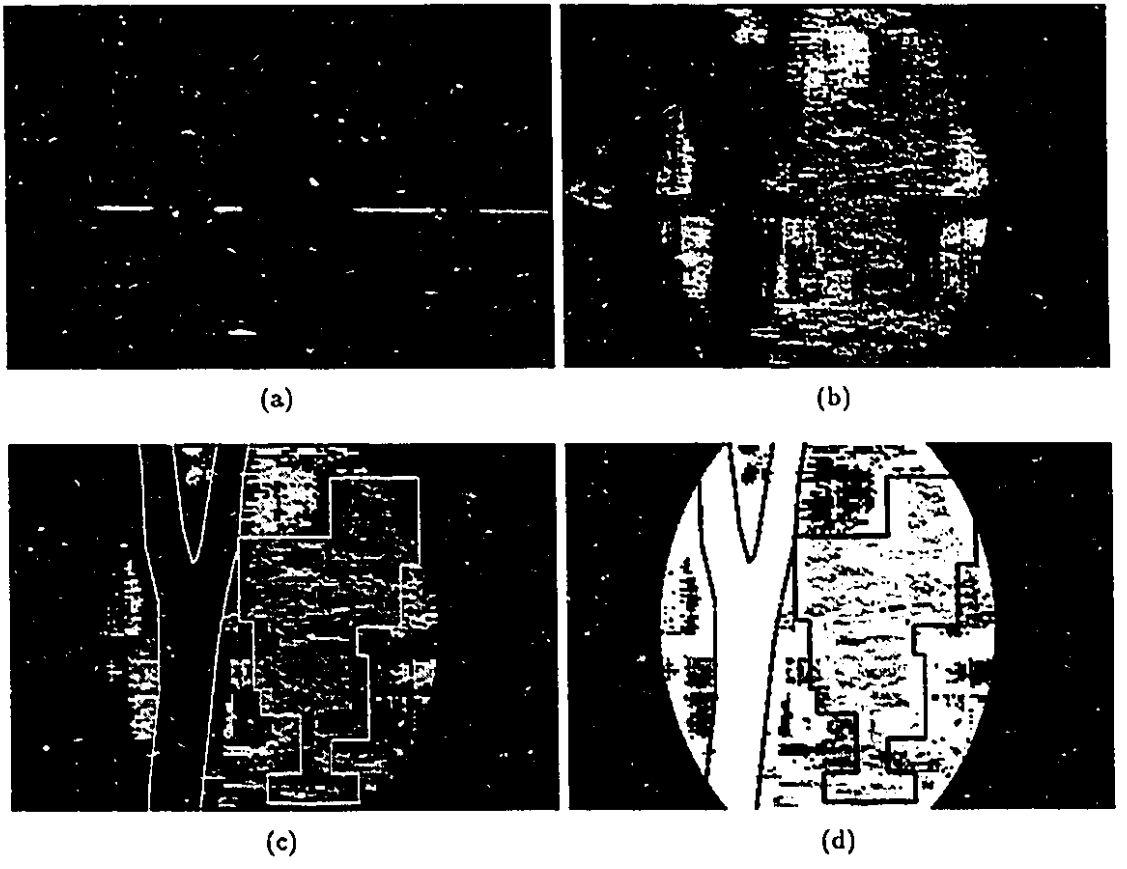


Figure 6.3: Scene of a tree trunk, sculpture and building. **(a)** A composite image of a scene consisting of (in depth order) a tree on the left, a sculpture on the right, and a building in the background. **(b)** The raw monocular disparity map given by cepstral analysis of the composite image in (a). Although the sculpture and building are at very different depths, they appear as very similar in disparity, since the camera is focused in front of the scene. **(c)** Reconstructed surfaces in disparity, given by maximum likelihood planar patches. The occluding boundaries of the tree and sculpture, as determined manually from the composite image, were superimposed in white. **(d)** Range image computed from the disparity map in (c). As in previous examples, the darker the intensity, the greater the depth. The difference in depth between the sculpture and background is now clear. The boundaries of the tree and sculpture were superimposed in black.

First, any areas of the composite image with little intensity variation over a horizontal extent larger than one window, are likely to contain many disparity errors. As described in Sec. 4.5, lack of image texture poses a problem similar to blur. There is not enough power across the Fourier spectrum with which to detect the ripple due to the visual echo. For example, on the face of the building in the background, there is a bright horizontal line from one side of the image to the other. In the raw disparity map, this region appears as erroneous disparities (dark), indicating that in the absence of image structure, smaller than expected echo delays were detected by the cepstrum. Similar difficulties occur in regions of the sculpture that are in shadow, and some areas of the background that are solid black. A second type of problem occurs at depth discontinuities in the scene, such as the occluding boundary of the tree and the sculpture. As illustrated in the experiments carried out in Sec. 5.3.2, when a cepstral window overlaps a depth discontinuity, a number of outcomes are possible. One surface may dominate over the other, so that in the disparity map objects seem to extend beyond their boundaries. This is the case in the fork of the tree. Another possibility is that the estimated disparity in these regions belongs to neither surface, as occurs along the lower right and upper left edges of the sculpture.

To generate a higher level representation of 3-D structure, local maximum likelihood planar patches were determined from the initial disparity estimates and the associated confidence values. If the goal is to localize step changes in disparity (obstacles), the size of these patches should be minimized, since any patch containing a discontinuity is likely to be unreliable. On the other hand, smaller patches provide less sample points on which to obtain a reliable estimate of local surface structure. One solution to this problem is to use a simpler local surface model, such as a fronto-parallel (one degree-of-freedom) planar patch, instead of the regular three degree-of-freedom patch. The results in Fig. 6.3c were obtained using 4×4 fronto-parallel patches. The boundaries of the tree and sculpture, determined manually from the composite image, were superimposed on the reconstructed surfaces for comparison. The result was converted into a range image (Fig. 6.3d), which clearly reveals not only the tree in the foreground, but also highlights the difference in depth between the sculpture

and the background.

Another example of the detection of a potentially hazardous obstacle is presented in Fig. 6.4a. Due to the low level of ambient light, vertical slit apertures were used instead of pinholes to allow more light to fall on the image plane. The camera was focused at a shallow depth, so that the background gave rise to relatively high disparities. Since monocular disparity and out-of-focus blur covary for non-pinhole apertures, in the composite image the background is very blurred (more so in the vertical direction than in the horizontal direction). Meanwhile, in the foreground, the toy Godzilla is both quite dark and contains relatively little intensity variation.

In the raw monocular disparity map (Fig. 6.4b), there are some errors in the background due to blur, and some errors in the foreground along the occluding contour of the toy. After the scene is reconstructed with 4×4 fronto-parallel planar patches, the obstacle is more clearly revealed in the foreground, yet there are sill errors in the background surface. Artifacts in the cepstrum introduced by the high degree of blur have likely caused some of the confidence estimates to be unreliable. This problem may be alleviated by fitting larger size facets to the background. The boundary of the obstacle, determined from manual inspection of the composite image, is superimposed in white (see Fig. 6.4c). When this result is converted to range (Fig. 6.4d), small errors in disparity on the background surface translate into larger errors in depth, due to the nonlinear relationship between disparity and depth. Nonetheless, despite the challenging nature of this scene, the obstacle in the foreground is detected and localized well enough that a mobile robot could maneouvre around it.

6.4 Locating Objects for Grasping

Another task for which range images are often used is to identify and locate objects at different depths so they may be acted upon by a machine. For example, suppose a camera is used to provide visual guidance for a robot arm and gripper. Presented with a collection of objects placed on a table-top, the robot is required to pick up a particular object with known dimensions. This requires not only calculating the

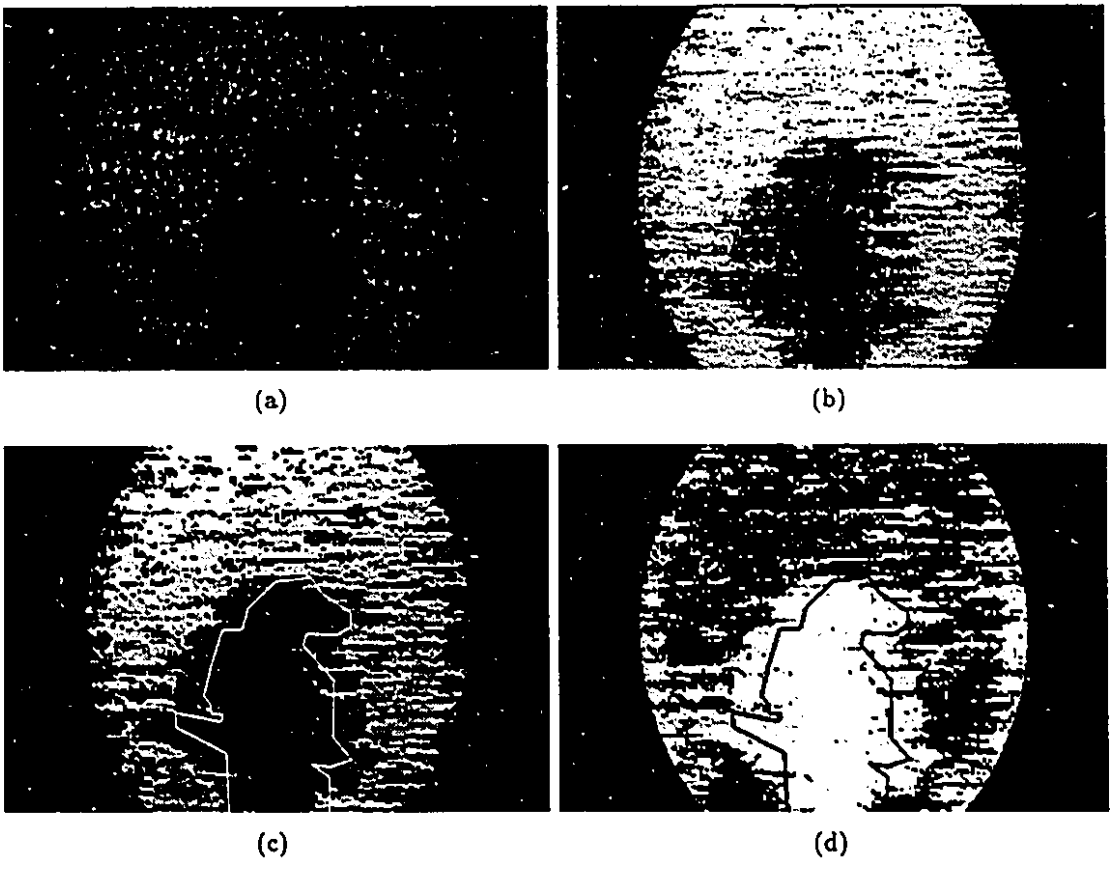


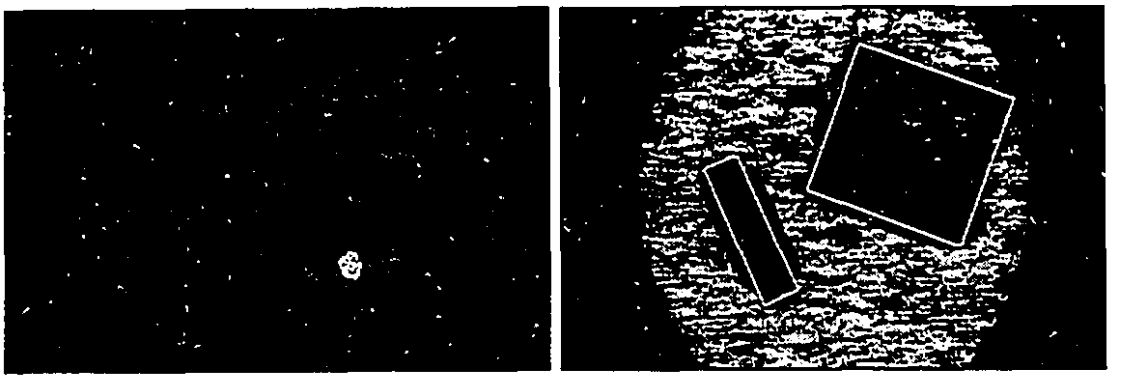
Figure 6.4: Scene of a toy Godzilla. **(a)** A composite image of a toy Godzilla monster placed in front of a textured background. Due to the low level of ambient level, vertical slit apertures were used instead if pinholes. Because of this, the background, where disparities are greatest, appears very blurred in the vertical direction, while less blurred in the horizontal direction. **(b)** Raw disparity map given by processing the composite image of (a) by the cepstral technique. Major errors occur in the background due to the high level of blur, and in the foreground due to lack of contrast. **(c)** Surface representation of the scene, given by fitting maximum likelihood 4×4 fronto-parallel patches to the raw disparity values. The outline of the toy monster was superimposed in white. **(d)** Range image given by converting (c) from disparity to depth. Darker grey levels correspond to greater depth. The outline of the toy monster is superimposed in black.

position and orientation of the desired object, but also determining a suitable path along which the robot arm can move without colliding into other objects or the table surface.

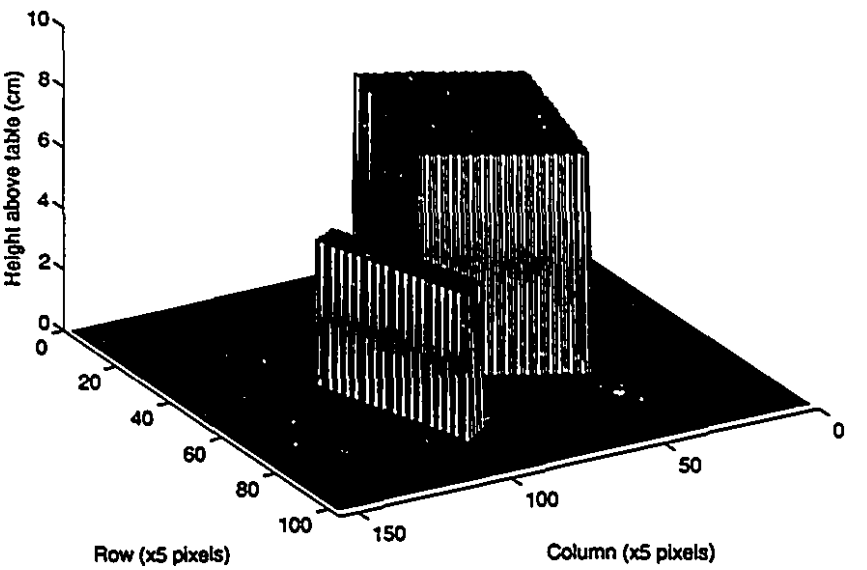
Consider the scene in Fig. 6.5 consisting of two objects on a table-top, as viewed above from a vertical angle. From this one image alone, even a human observer is unable to judge the relative depth of the two objects. In order to pick up one object without knocking down the other, the height of both objects is required. For example, if the desired object is 15 cm in height and the other 20 cm, the robot must be careful not to collide with the taller object, while approaching the desired object with its gripper. On the other hand, if the second object is much smaller, say, 2 cm in height, the robot is free to operate in the space above the smaller object.

In the raw disparity map determined by cepstral analysis of the composite image (Fig. 6.5b), the two objects are clearly detected, but in the background, where texture is sparse and blur is more significant, there are more noticeable errors. Rather than reconstruct the scene with planar facets, a more sophisticated model of the environment can be exploited to obtain better results. For example, the scene can be modelled as several planar objects placed on a fronto-parallel plane. The original intensity image or raw disparity map can be segmented to identify regions of the composite image corresponding to these objects. For example, suppose the segmentation process identified the polygonal regions outlined in white in Fig. 6.5b as two objects standing on the table-top. A single fronto-parallel planar surface was fit (in a maximum likelihood framework) to each segmented region, and to the background region representing the table surface. Knowing that the objects are closer to the camera than the table-top, any disparity estimates in the background region that correspond to a depth less than the objects, are assigned a probability correct of zero. When the resulting surface representation is converted from disparity to depth and displayed as a mesh plot, the relative depth of the two objects is clearly evident (see Fig. 6.5c).

For a second example, four objects placed on a table-top were viewed with a CCD camera with twin pinhole apertures (see Fig. 6.6a). There are no areas at the left and right edges of the composite image where the view from only one aperture is



(a) (b)



(c)

Figure 6.5: Scene of two objects on a table-top. **(a)** A composite image of a two objects placed on a fronto-parallel planar surface (a table). Due to the low level of ambient level, vertical slit apertures were used instead if pinholes. Because of this, the background, where disparities are greatest, appears very blurred in the vertical direction, while less blurred in the horizontal direction. **(b)** Raw disparity map given by processing the composite image of (a) by the cepstral technique. Disparity errors occur in the background primarily in those areas containing very few dots (i.e., insufficient texture). **(c)** Assuming the segmentation of the disparity map indicated by the white polygons in (b), a ML fronto-parallel plane was fit to each of the four objects and the ground plane, converted to depth, and displayed as a mesh plot.

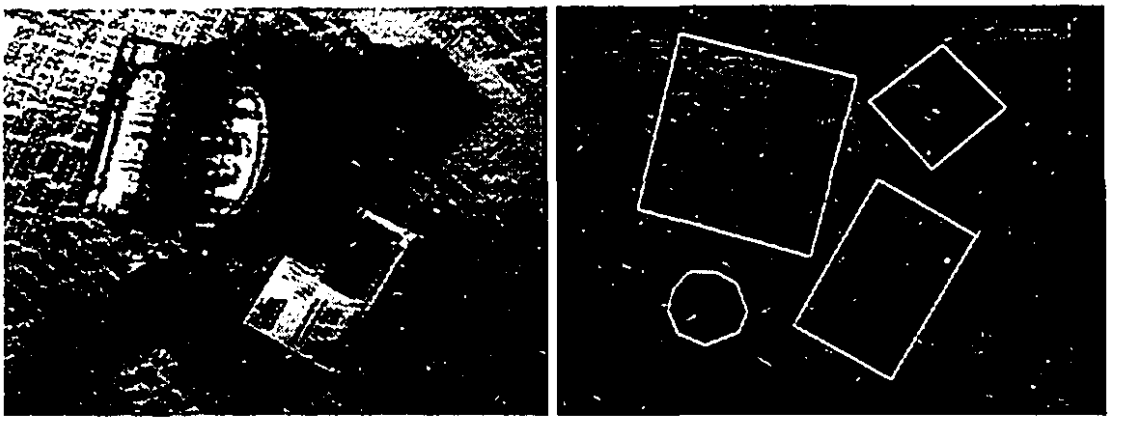
visible, due to the different configuration of the double aperture CCD camera. In the raw disparity map there is a high density of noticeable errors in the four corners (see Fig. 6.6b). In these regions the visual echo is not perfectly horizontal because of lens distortion, so a 1-D cepstrum taken along a scanline is often unable to detect the visual echo. To address this problem, 2-D cepstra could be computed to enable detection of a visual echo in any orientation, not just horizontal. Another effect of lens distortion causes disparity of a fronto-parallel plane to vary over a wide region. Instead of looking like a plane, the surface in disparity space is slightly bowl shaped, the lowest point being in the centre of the composite image. By calibrating the camera, this effect can be measured and removed from the raw disparity map.

Assuming the raw disparity map is segmented into the four polygonal regions outlined in Fig. 6.6b, the scene model of planar objects standing on a ground plane can again be exploited to obtain good results (see Fig. 6.6c). Despite the close proximity of the four objects and lower resolution of the composite image, the objects are sufficiently localized for a robot grasping task.

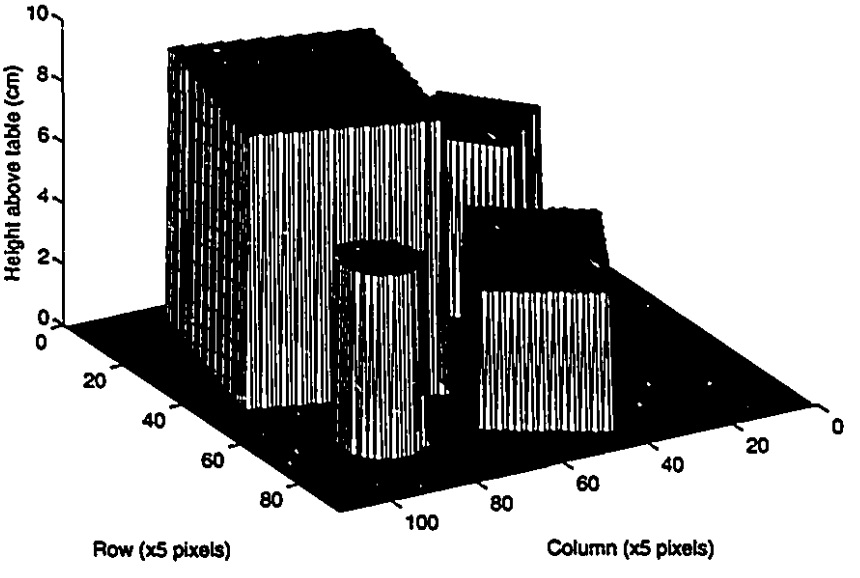
6.5 Robot Navigation

The final experiment is meant to demonstrate how monocular stereopsis can be used to guide an autonomous machine through an unknown, unstructured environment. The particular environment consists of a lounge area containing cabinets, bookshelves, chairs, and tables (see Fig. 6.7). Composite images were acquired with the SLR camera from different viewpoints in this room, and converted, using the technique developed in this thesis, into a 3-D representation of surfaces in the scene. This representation was converted back into a range image for interpretation and display purposes. All the images were taken with the camera at the same height, aligned so that both the apertures and the image scanlines were parallel to the ground plane. These images simulate the views seen by a mobile robot with one double aperture camera, mounted in a fixed position on top of the robot.

The goal in the following discussion is to illustrate how a mobile robot can use



(a) (b)



(c)

Figure 6.6: Scene of four objects on a table-top. **(a)** A composite image of four objects placed on a fronto-parallel planar surface (a table). This image was taken with a double aperture CCD camera. **(b)** Raw disparity map given by processing the composite image of (a) by the cepstral technique. **(c)** Assuming the segmentation of the disparity map indicated by the white polygons in (b), a ML fronto-parallel plane was fit to each of the four objects and the ground plane, converted to depth, and displayed as a mesh plot.

the range images provided by monocular stereopsis to accomplish a required task within the lounge area. In particular, this "simulated robot" is required to navigate from a starting position (in the lower left-hand corner), to its destination (in the upper right-hand corner), where a second robot is in need of repair and emitting a beacon (or perhaps it is just lost and crying for help). It is assumed that a double aperture camera is mounted on the robot in a fixed position, so that it can look in only one direction, that which the robot considers to be "straight ahead". This viewing direction is indicated in Figs. 6.7 and 6.9 by a solid arrow originating from the robot position. Besides the beacon emitted by the defective robot, the only information the mobile robot has about its environment is what it can obtain from monocular stereopsis. Based on this information, the robot must get to its defective partner without colliding into any furniture or walls.

From its starting point, the robot views a scene consisting of chair A in the foreground on the right, and on the left, an open space all the way back to the bookshelf (see Fig. 6.8a). The composite image was processed by the cepstrum yielding a raw disparity map (Fig. 6.8b). The characteristics and processing parameters for all the composite images in this section are given in Tables 6.1 and 6.2. Surfaces in the scene were reconstructed by fitting 16×16 maximum likelihood planar patches (Fig. 6.8c) to the raw disparity map. Finally, these surfaces were converted from disparity space to 3-D space, and displayed as a range image (Fig. 6.8d). This surface representation is quite good despite errors in the raw disparity map, indicating that the confidence measure has correctly labelled low confidence disparity estimates. For display purposes, in the range image, darker grey level intensities correspond to greater depth in the scene.

Based on this range image, the robot can conclude two important facts. First, if it drives straight ahead it will collide with a large obstacle approximately 0.8 m away (chair A). Second, to the left of this obstacle is an area of free space extending for some 3 m. Assuming the robot has some "intelligence", it will therefore decide to turn to the left, into this zone of free space. The actual process by which this decision is made, in the context of robot path planning, is outside the scope of this

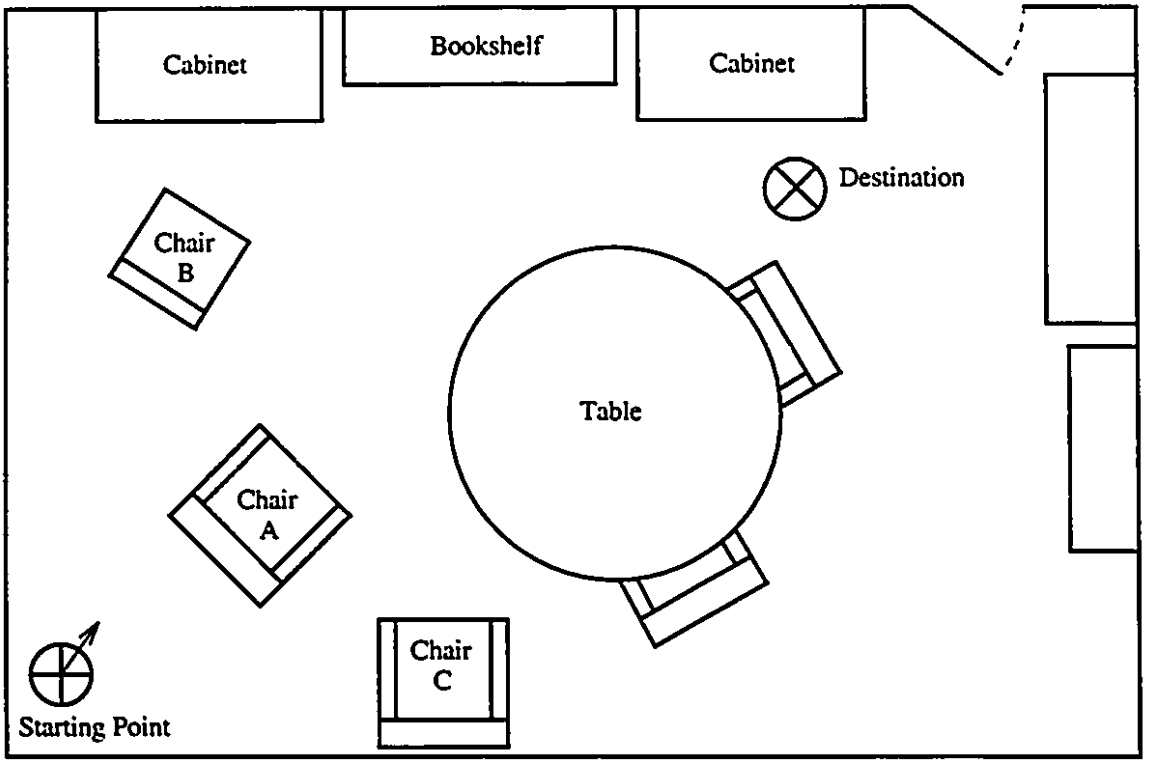


Figure 6.7: Map of lounge area to be navigated by mobile robot. A mobile robot is required to navigate a path from its starting point in the lower left, to its destination in the upper right. Based on the range images provided by a multiple aperture camera, the robot must avoid colliding with any of the furniture in the room.

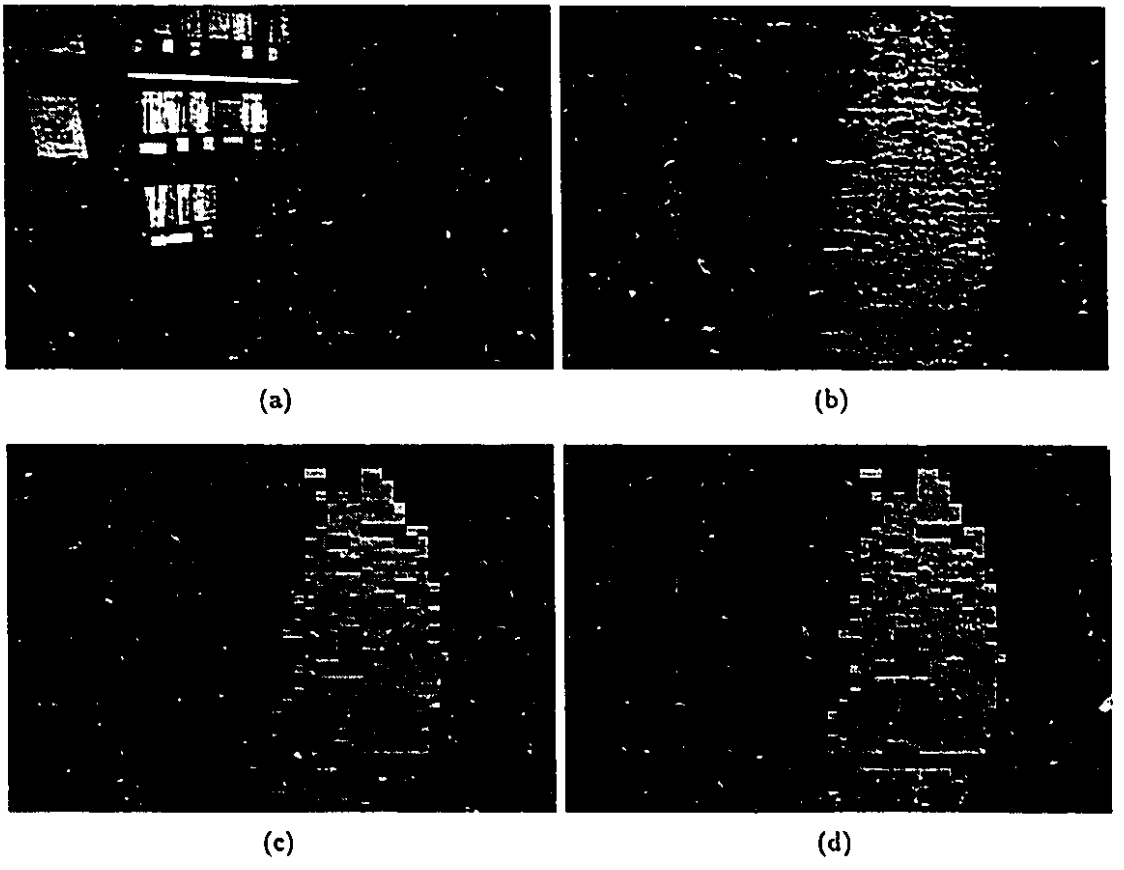


Figure 6.8: Scene from robot position 1. **(a)** The composite image of the lounge area as seen from the initial position and orientation of the mobile robot. This particular image was taken with the 35 mm SLR camera, with two vertical slit apertures, digitized at 1536 × 1024 8-bit resolution. **(b)** Raw disparity map given by cepstral analysis of 128×16 sliding windows applied with a step size of 2 × 2 pixels. Due to the periodic texture on the back of the chair on the right, and the horizontal areas of constant intensity in the background, the raw disparity map contains some noticeable errors. **(c)** Reconstructed disparity map, given by the maximum likelihood 16 × 16 local planar patches determined from the raw disparity map and the associated confidence values. Most of the major errors in (b) have been implicitly identified and removed. **(d)** The range image given by converting (c) from monocular disparity to depth. In this image, the darker a pixel intensity, the greater the distance in the scene to that point.

thesis. For example, the robot could rotate in its current position to further explore its environment, before proceeding. Similar range images acquired at fixed rotational increments, would reveal the rest of chair A, chair C, the walls behind the robot, and chair B. By integrating range information acquired from these multiple viewpoints, the robot could begin to construct a map of its environment similar to the map in Fig. 6.7. Assuming some *a priori* knowledge of the approximate location of its destination (such as a beacon from a defective robot), or alternatively, a strategy to explore and identify its target, the conclusion would be to proceed into the free space to the left of chair A. The goal here is not to explain the details of robot navigation or map building, but to show how monocular stereopsis provides the sensory information required to complete these tasks.

After moving to the left of chair A, the robot must consider its next move. From this second viewpoint (see Fig. 6.9), the robot is able to detect chair B on the left and free space on the right (Fig. 6.10). Therefore it decides to turn right, avoiding a collision with chair B. From the third viewpoint, the robot sees only the bookcase in the distance (Fig. 6.11), and therefore decides to continue on its current course. From position 4, the bookcase is now very close (Fig. 6.12), so the robot must turn again to avoid a collision. Finally, from position 5, a clear path is seen to the required destination (Fig. 6.13).

The range information acquired from these five viewpoints may be integrated to form a crude map of the robot's environment. To achieve this, each range image was processed as follows. First, the range image was converted into a set of 3-D surface points, expressed in a coordinate frame given by the position and orientation of the camera when the composite image was acquired. To display the final data in a more compact format, each column of the range image was divided into 20 equal-length segments, and the median depth value in each segment recorded. To display the final data as if the room were viewed from above (i.e., a map), the Y-coordinates (corresponding to position along the columns of the range image) were discarded. The resulting data was then transformed into a common, global coordinate system and combined with data from other viewpoints. The global coordinate space, as viewed

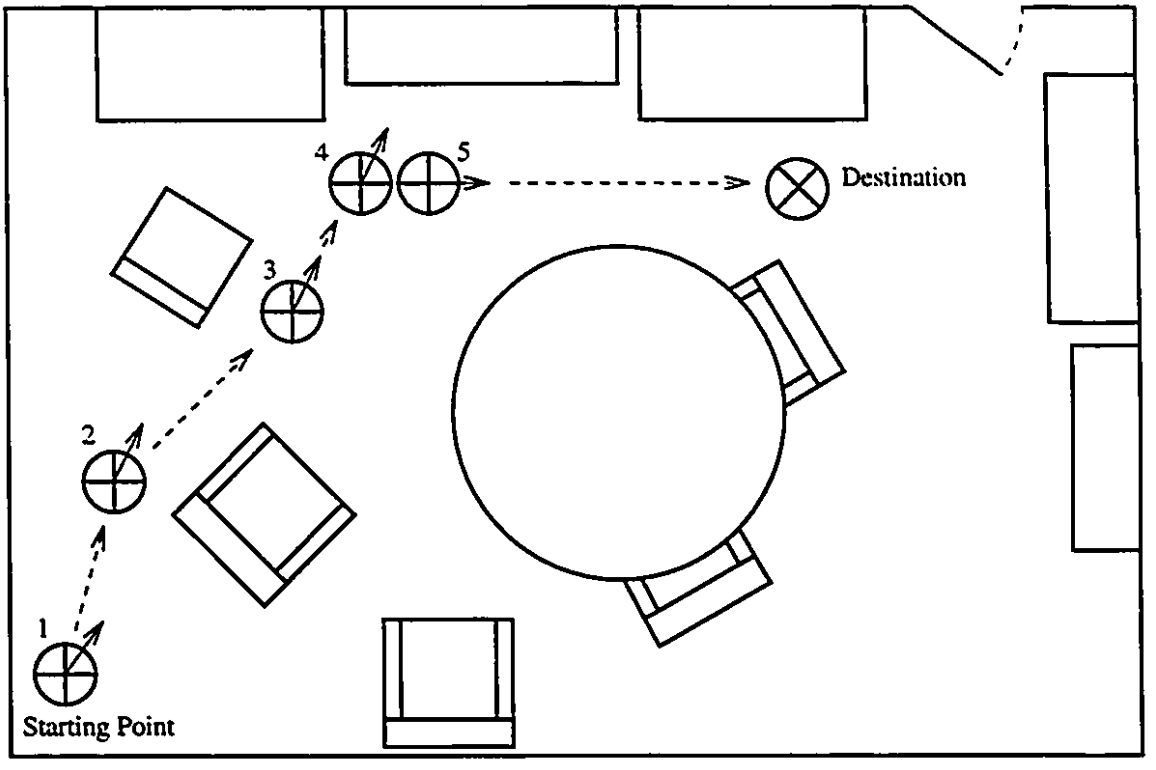


Figure 6.9: Path taken by robot from starting point to destination. At each of the five positions labelled, a composite image of the scene was acquired (in the direction indicated by the solid line arrows) and converted into a range image as described in the text. Based the range image from a given position, the mobile robot can determine its next move, indicated by the dashed line arrows.

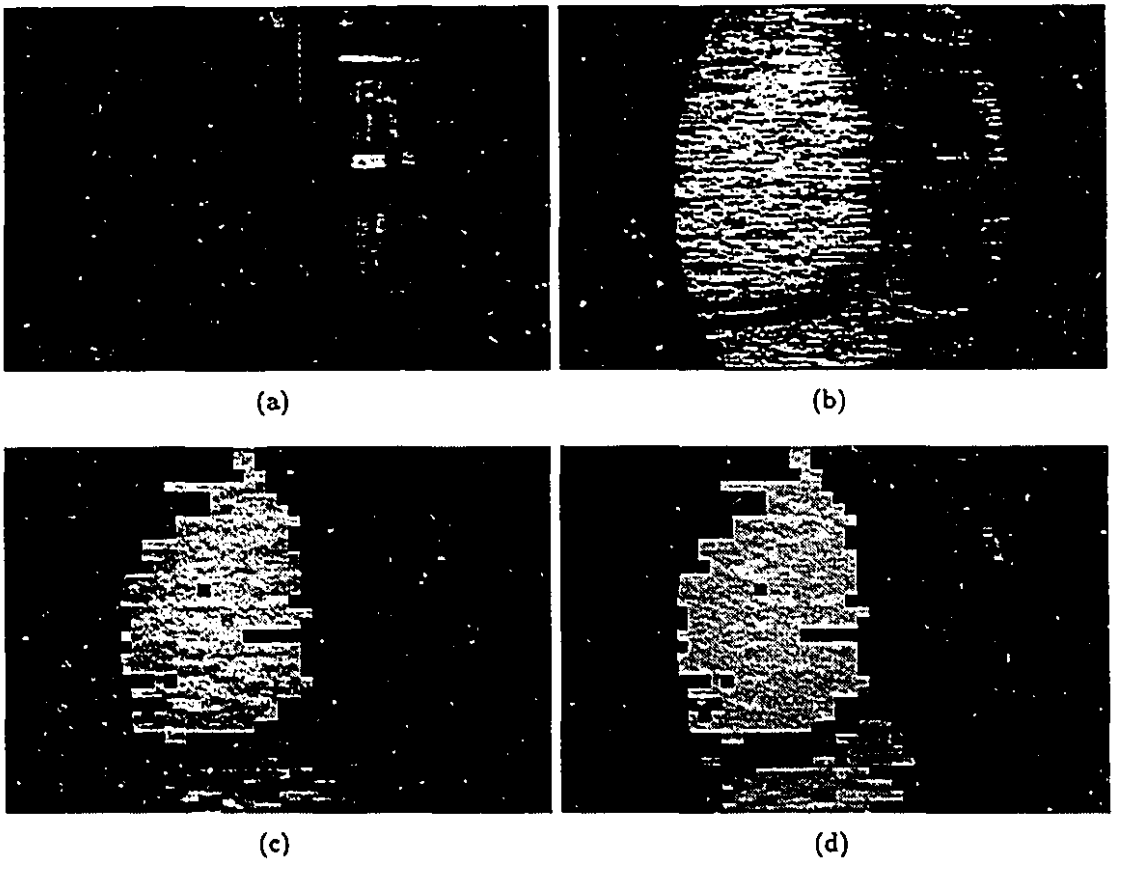


Figure 6.10: Scene from robot position 2. **(a)** Composite image taken from robot position 2 in Fig. 6.9. **(b)** Raw disparity map given by cepstral analysis. **(c)** Reconstructed surfaces given by maximum likelihood local planar patches. **(d)** Range image, where depth is displayed as a grey level intensity according to the same scale as in Fig. 6.8d.

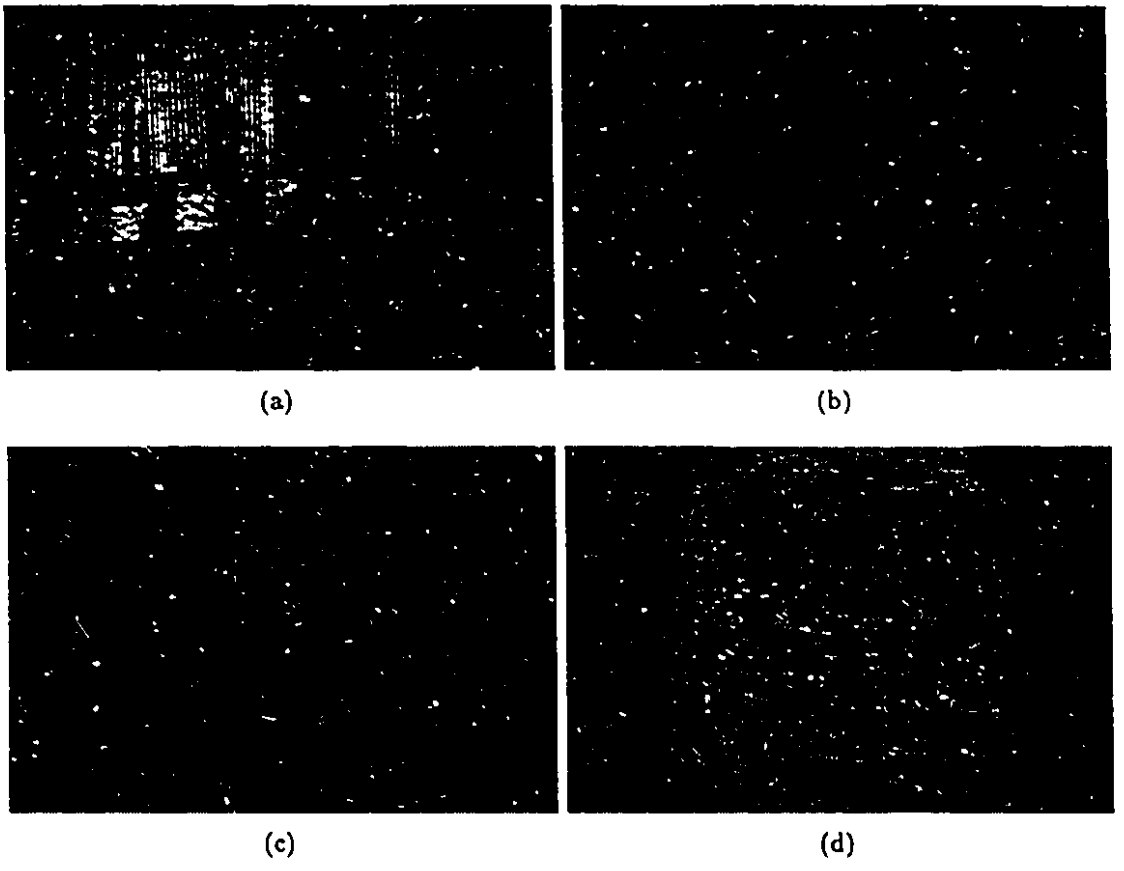


Figure 6.11: Scene from robot position 3. **(a)** Composite image taken from robot position 3 in Fig. 6.9. **(b)** Raw disparity map given by cepstral analysis. **(c)** Reconstructed surfaces given by maximum likelihood local planar patches. **(d)** Range image, where depth is displayed as a grey level intensity according to the same scale as in the previous figures.

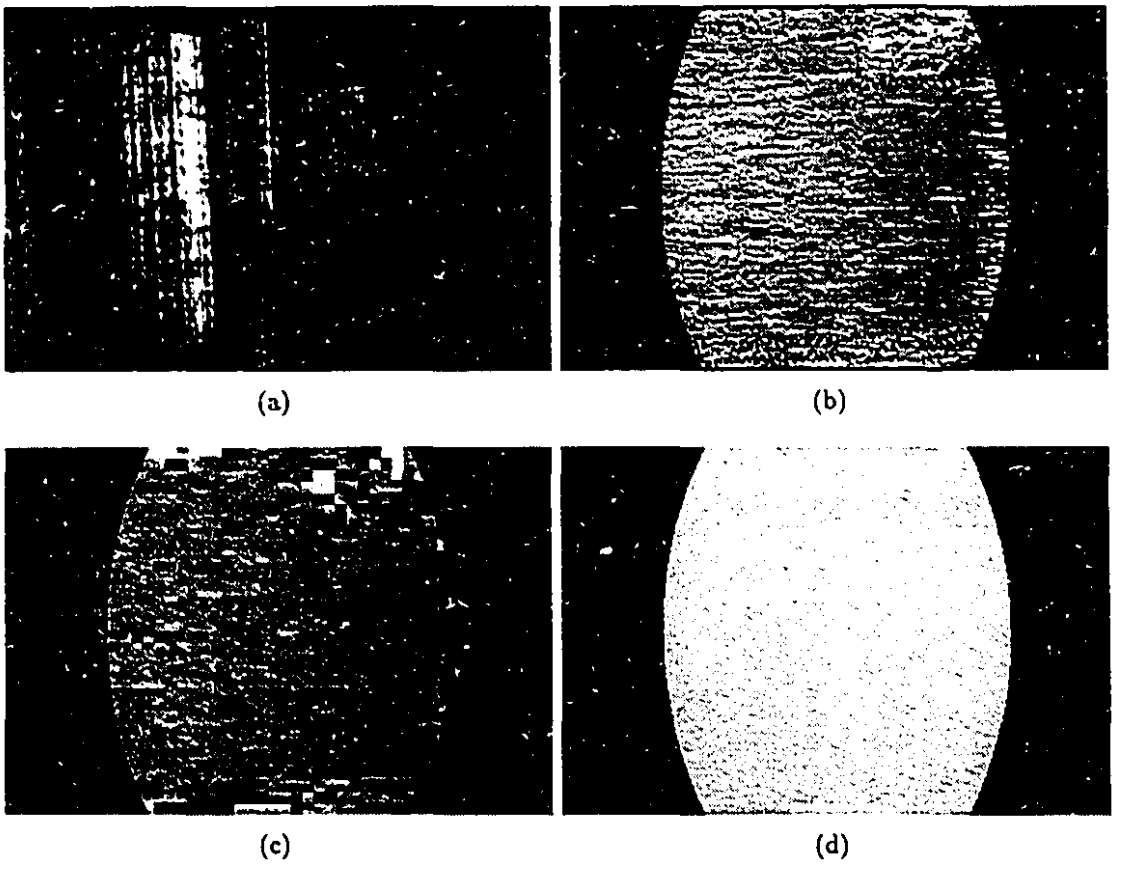


Figure 6.12: Scene from robot position 4. **(a)** Composite image taken from robot position 4 in Fig. 6.9. **(b)** Raw disparity map given by cepstral analysis. **(c)** Reconstructed surfaces given by maximum likelihood local planar patches. **(d)** Range image, where depth is displayed as a grey level intensity according to the same scale as in the previous figures.

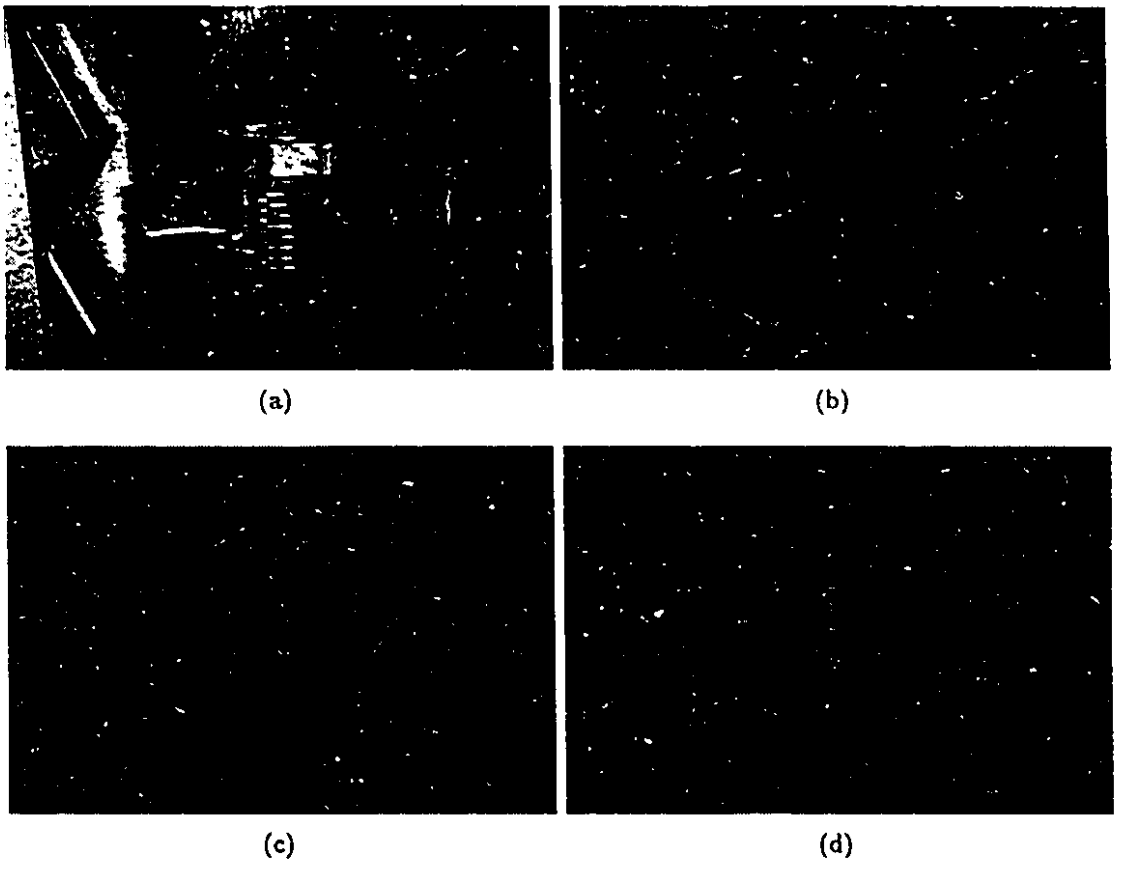


Figure 6.13: Scene from robot position 5. **(a)** Composite image taken from robot position 5 in Fig. 6.9. **(b)** Raw disparity map given by cepstral analysis. **(c)** Reconstructed surfaces given by maximum likelihood local planar patches. **(d)** Range image, where depth is displayed as a grey level intensity according to the same scale as in the previous figures.

from above like a map, was divided into a fine 2-D grid. The integrated range data were assigned to the cells of this grid, such that a counter in each cell was incremented each time a data point fell into that cell.

The resulting grid is displayed as an intensity image, superimposed on the actual map of the lounge (at the same scale) in Fig. 6.14. The darker the intensity at a given position in the map, the higher the density of range data occurring at that position. The dashed lines emerging from each viewpoint position indicate the usable field of view in the composite image. Notice that the closer a surface point to the viewpoint, the more accurately its depth is measured. Sub-pixel disparity errors of comparable size correspond to small depth uncertainty at near viewing distances and much greater depth uncertainty at farther viewing distances (see Fig. 3.2). For example, from viewpoint 5, the defective robot and its open panel are well localized, while there is much more scatter in the data around the cabinet in the background. Small errors in measuring the disparity of the cabinet translate into large errors in depth.

The construction of an accurate map of the environment for the purposes of robot navigation has been dealt with in detail elsewhere (e.g., using dense sonar range data [49]). Clearly five scenes is not enough to determine a complete map, but this exercise shows that one multiple-aperture camera provides range data of sufficient resolution and accuracy that, given enough viewpoints, such a map could be computed. It is worthwhile noting that while the map of the lounge constructed above is incomplete, it was sufficient for the robot to achieve the required task. Furthermore, this technique is passive, truly monocular, and can be implemented in hardware, to provide an inexpensive, real-time range sensor.

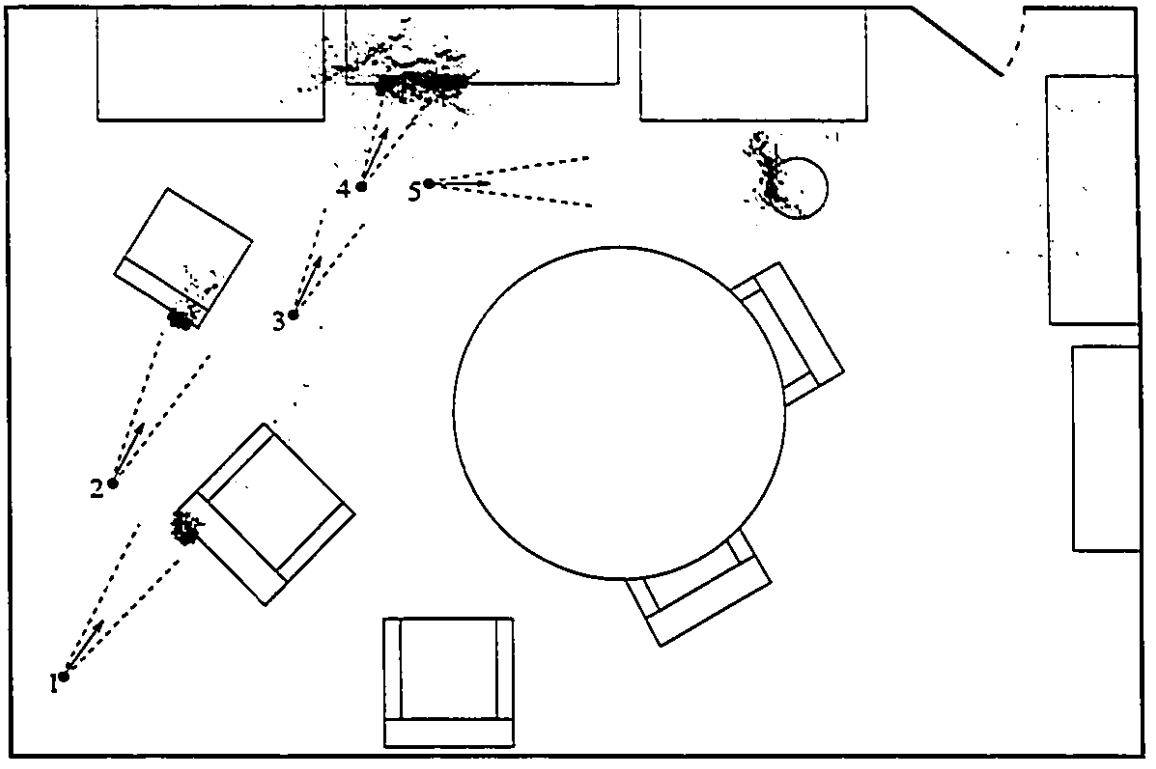


Figure 6.14: Map constructed by integrating range data from five views. Range data from the five viewpoints, converted into a common global coordinate frame, is displayed as a grey level image superimposed (at the same scale) onto an actual map of the environment. The darkness of the data plotted corresponds to the relative frequency of range data points in that region of space. The dotted lines emerging from each viewpoint position indicate the usable field of view in each composite image.

Chapter 7

Conclusions

A multiple aperture camera may be used to compute an accurate range image from one composite image. Depth is encoded by the displacement or disparity between points on the image plane projecting from the same point in the scene. Unlike binocular stereopsis, in *monocular stereopsis* eye of origin information is lost, therefore conventional solutions to the binocular correspondence problem are unable to measure monocular disparity.

Cepstral analysis offers a solution to this problem. The cepstrum of a composite image window exhibits a peak at the monocular disparity value. The proposed model of the composite image cepstrum predicts both the shape and height of this peak, and the nature of the noise in the cepstrum that may obscure this peak. This leads to a two-stage algorithm for measuring monocular disparity to sub-pixel precision. Associated with each stage is a confidence measure that predicts the distribution of measurement error. The weighted combination of these two distributions provides an overall probability density function for each disparity measurement. This density function, combined with some local surface model, allows a maximum likelihood reconstruction of surfaces in the scene.

It is inherent to the cepstral technique that disparity estimates are made over a composite image window rather than at a single pixel. This would seem to suggest that obstacles of width less than the window size may not be detectable, and edges in

depth are poorly localized. This is not the case. Windows may be centered on every pixel of the composite image, the estimated disparity over each window recorded at the centre pixel. When the resulting disparity map and confidence values are analyzed in a maximum likelihood framework, obstacles as narrow as one-eighth the window width may be reliably detected, and depth edges may be localized to within one-eighth the window width. This provides adequate spatial resolution for many applications of range imaging.

In terms of computation, the techniques described in this thesis are relatively straightforward. Measurement of disparity by cepstral analysis involves two FFT (or FHT) operations and a logarithm, a peak detection, and the evaluation of some simple expressions for sub-pixel disparity localization. Calculation of the confidence measure involves a few more simple expressions and several table lookups. Furthermore, composite image windows may be processed completely independently. Taken together, these observations imply that this technique is suitable for parallel implementation in hardware, providing a range sensor that may truly operate in real-time. Even the surface reconstruction procedure, often considered a computationally expensive task, can be implemented in parallel or in hardware.

The experimental results presented illustrate how this range sensor may be used for tasks such as mobile robot navigation and collision avoidance. Compared to binocular stereo range sensors, the proposed sensor is less expensive, more compact, requires only one video channel, and can be implemented in real-time. With these practical advantages in mind, passive monocular range imaging with a multiple aperture camera should be considered as a possible solution to many problems requiring the automated recovery of 3-D scene structure.

Appendix A

Planar Facets in Disparity Space

The relationship between monocular disparity and depth is nonlinear. Therefore an object of some shape in 3-D space may correspond to a quite different shape in disparity space. It is often appropriate to approximate surfaces in 3-D space as being locally planar. Through analysis of the equations relating image coordinates, 3-D coordinates, and monocular disparity, it is possible to show that a plane in depth corresponds to a plane in disparity.

Let (X, Y, Z) be a world coordinate system with origin at the centre of the image plane and Z-axis corresponding to the optical axis of a double aperture camera. Define the camera to have focal length F, effective aperture separation D, distance from the lens to the sensor plane f, and the two apertures to lie on the X-axis equally spaced about the origin. Assume the camera is focused at a depth of infinity, so that all monocular disparities are positive. Let $P(X_w, Y_w, Z_w)$ be a point in the scene, which when projected through each aperture gives rise to points $P_1(x, y)$ and $P_2(x + d_P, y)$ on the image plane, where d_P is the monocular disparity value. The coordinates of P are therefore given by the following equations

$$\frac{1}{Z_w} = \frac{1}{F} - \frac{1}{f}\left(1 - \frac{d_P}{D}\right) \tag{A.1a}$$

$$X_w = \frac{Z_w\,(x + d_P/2)}{f} \tag{A.1b}$$

$$Y_w = \frac{Z_w y}{f} \tag{A.1c}$$

as developed in Eqns. (3.4b), (6.1a) and (6.1b). Solving Eqn. (A.1a) for Z_w, and substituting the result into Eqns. (A.1b) and (A.1c), gives

$$Z_w = \frac{DfF}{Df - F(D + d_P)} \tag{A.2a}$$

$$X_w = \frac{DF(x + d_P/2)}{Df - F(D + d_P)} \tag{A.2b}$$

$$Y_w = \frac{DFy}{Df - F(D + d_P)} \tag{A.2c}$$

If P lies on a plane, the following relationship exists between its coordinates

$$Z_w = A X_w + B Y_w + C \tag{A.3}$$

where A, B, C are the parameters of the plane in 3-D space. Substituting Eqns. (A.2a), (A.2b), and (A.2c) into Eqn. (A.3) and solving for d_P gives

$$\begin{aligned} d_P &= \frac{2ADFx + 2BDFy - 2D(fF - Cf + CF)}{2CF - ADF} \\ &= A'x + B'y + C' \end{aligned} \tag{A.4}$$

where A', B', C' are constants. Therefore for any scene point lying on a plane in 3-D space (satisfying Eqn. (A.3)), the corresponding disparity value lies on a plane in image space (satisfies Eqn. (A.4)). Hence a plane in depth corresponds to a plane in monocular disparity.

References

[1] E.H. Adelson and J.Y.A. Wang, "Single lens stereo with a plenoptic camera", *IEEE Trans. Pattern Analysis and Machine Intelligence* 14(2):99–106, 1992.

[2] N. Ayache and B. Faverjon, "Efficient registration of stereo images by matching graph descriptions of edge segments", *International Journal of Computer Vision* 1:107–131, 1987.

[3] H.H. Baker and T.O. Binford, "Depth from edge and intensity based stereo", in *Proc. 7th International Joint Conf. on Artificial Intelligence*, pp. 631–636, 1981.

[4] E. Bandari and J.J. Little, "Cepstral analysis of optical flow". Technical Report 92-6, Dept. of Computer Science, Univ. of British Columbia, 1992.

[5] E. Bandari and J.J. Little, "Multi-evidential correlation and visual echo analysis". Technical Report 93-1, Dept. of Computer Science, Univ. of British Columbia, 1993.

[6] E. Bandari and J.J. Little, "Visual echo analysis", in *Proc. 4th International Conference on Computer Vision*, pp. 220–225, 1993.

[7] S.T. Barnard, "A stochastic approach to stereo vision", in *Proc. 5th National Conf. on Artificial Intelligence*, pp. 676–680, 1986.

[8] S.T. Barnard and W.B. Thompson, "Disparity analysis of images", *IEEE Trans. Pattern Analysis and Machine Intelligence* 2(4):333–340, 1980.

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

[10] F. Blais and M. Rioux, "BIRIS: a simple 3-D sensor", in *Proc. SPIE Vol. 728 Optics, Illumination, and Image Sensing for Machine Vision*, pp. 235–242, 1986.

[11] A. Blake and A. Zisserman, *Visual Reconstruction*. MIT Press, 1987.

[12] B.P. Bogert, M.J.R. Healy, and J.W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking", in *Proc. Symposium on Time Series Analysis*, pp. 209–243, 1963.

[13] B.P. Bogert and J.F. Ossanna, "The heuristics of cepstrum analysis of a stationary complex echoed gaussian signal in stationary gaussian noise", *IEEE Trans. Information Theory* 12(3):373–380, 1966.

[14] V.M. Bove, "Entropy-based depth from focus", *J. Optical Society of America A* 10(4):561–566, 1993.

[15] R.N. Bracewell, *The Hartley Transform*. Oxford University Press, 1986.

[16] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina", *Proc. Royal Society London B* 220:89–113, 1983.

[17] D.G. Childers, D.P. Skinner, and R.C. Kemerait, "The cepstrum: a guide to processing", *Proc. of the IEEE* 65(10):1428–1443, 1977.

[18] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., 1973.

[19] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus", *IEEE Trans. Pattern Analysis and Machine Intelligence* 15(2):97–107, 1993.

[20] R. Fabian and D. Malah, "Robust identification of motion and out-of-focus blur parameters from blurred and noisy images", *CVGIP: Graphical Models and Image Processing* 53(5):403–412, 1991.

[21] D.J. Field, "Relations between the statistics of natural images and the response properties of cortical cells", *J. Optical Society of America A* 4(12):2379–2394, 1987.

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

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

[24] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Analysis and Machine Intelligence* 6(6):721–741, 1984.

[25] D.B. Gennery, "A stereo vision system for an autonomous vehicle", in *Proc. 6th International Joint Conf. on Artificial Intelligence*, pp. 576–582, 1979.

[26] A. Goshtasby and W.A. Gruver, "Design of a single-lens stereo camera system", *Pattern Recognition* 26:923–937, 1993.

[27] D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics*. John Wiley and Sons, Inc., 1966.

[28] W.E.L. Grimson, "A computer implementation of a theory of human stereo vision", *Phil. Trans. Royal Society of London* B292:217–253, 1981.

[29] W.E.L. Grimson, "An implementation of a computational theory of visual surface interpolation", *Computer Vision, Graphics and Image Processing* 22:39–69, 1983.

[30] W.E.L. Grimson and T. Pavlidis, "Discontinuity detection for visual surface reconstruction", *Computer Vision, Graphics and Image Processing* 30:316–330, 1985.

[31] P. Grossmann, "Depth from focus", *Pattern Recognition Letters* 5:63–69, 1987.

[32] J.C. Hassab and R. Boucher, "A probabilistic analysis of time delay extraction by the cepstrum in stationary Gaussian noise", *IEEE Trans. Information Theory* 22(4):444–454, 1976.

[33] J.C. Hassab and R. Boucher, "Improved cepstrum performance through windowing of log spectrum", *J. Sound and Vibration* 58(4):597–598, 1978.

[34] E. Hecht, *Optics*. Addison-Wesley, 2nd edition, 1987.

[35] R.A. Hummel and S.W. Zucker, "On the foundations of relaxation labeling processes", *IEEE Trans. Pattern Analysis and Machine Intelligence* 5(3):267–287, 1983.

[36] B. Jahne and P. Geissler, "Depth from focus with one image", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 713–717, 1994.

[37] A.D. Jepson and M.R.M. Jenkin, "The fast computation of disparity from phase differences", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 398–403, 1989.

[38] D.G. Jones, *Computational Models of Binocular Vision*. PhD thesis, Stanford University, 1991.

[39] D.G. Jones and D.G. Lamb, "Analyzing the visual echo: Passive 3-D imaging with a multiple aperture camera". Technical Report CIM-93-3, McGill Research Centre for Intelligent Machines, 1993.

[40] D.G. Jones and J. Malik, "A computational framework for determining stereo correspondence from a set of linear spatial filters", in *Proc. 2nd European Conference on Computer Vision*, pp. 395–410, 1992.

[41] D.G. Jones and J. Malik, "Determining three-dimensional shape from orientation and spatial frequency disparities", in *Proc. 2nd European Conference on Computer Vision*, pp. 661–669, 1992.

[42] B. Julesz, "Binocular depth perception of computer generated patterns", *Bell Systems Technical Journal* 39:1125–1162, 1960.

[43] T. Kanungo, M.Y. Jaisimha, J. Palmer, and R.M. Haralick, "A quantitative methodology for analyzing the performance of detection algorithms", in *Proc. 4th International Conference on Computer Vision*, pp. 247–252, 1993.

[44] R.C. Kemerait and D.G. Childers, "Signal detection and extraction by cepstrum techniques", *IEEE Trans. Information Theory* 18(6):745–759, 1972.

[45] Y.C. Kim and J.K. Aggarwal, "Positioning three-dimensional objects using stereo images", *IEEE J. Robotics and Automation* 3(4):361-373, 1987.

[46] E. Krotkov, "Focusing", *International Journal of Computer Vision* 1:223–237, 1987.

[47] S.-H. Lai, C.-W. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image", *IEEE Trans. Pattern Analysis and Machine Intelligence* 14(4):405–411, 1992.

[48] K.-O. Ludwig, H. Neumann, and B. Neumann, "Local stereoscopic depth estimation using ocular stripe maps", in *Proc. 2nd European Conference on Computer Vision*, pp. 373–377, 1992.

[49] P. Mackenzie. "Mobile Robot Localization Using Model-Based Maps". Master's thesis, McGill University, Dept. of Electrical Engineering, 1994.

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

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

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

[53] P. Meer, D. Mintz, A. Rosenfeld, and D.Y. Kim, "Robust regression methods for computer vision: A review", *International Journal of Computer Vision* 6:59–70, 1991.

[54] H.P. Moravec, "Rover vehicle obstacle avoidance", in *Proc. 7th International Joint Conf. on Artificial Intelligence*, pp. 785–790, 1981.

[55] S.K. Nayar, "Shape from focus system", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 302-308, 1992.

[56] A.M. Noll, "Cepstrum pitch determination", *J. Acoustical Society of America* 41(2):293-309, 1966.

[57] Y.O. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming", *IEEE Trans. Pattern Analysis and Machine Intelligence* 7(2):139-154, 1985.

[58] T.J. Olson and D.J. Coombs, "Real-time vergence control for binocular robots", *International Journal of Computer Vision* 5:67-89, 1991.

[59] A.V. Oppenheim, R.W. Schafer, and T.G. Stockham, "Nonlinear filtering of multiplied and convolved signals", *Proc. of the IEEE* 56(8):1264-1291, 1968.

[60] A. Pentland, T. Darrell, M. Turk, and W. Huang, "A simple real-time range camera", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 256-261, 1989.

[61] A.P. Pentland, "A new sense for depth of field", *IEEE Trans. Pattern Analysis and Machine Intelligence* 9(4):523-531, 1987.

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

[63] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C - The Art of Scientific Computing*. Cambridge University Press, 1988.

[64] M. Rioux. "Three dimensional imaging device". U.S. Patent 4,645,347, 1987.

[65] M. Rioux. "Three dimensional imaging device comprising a lens system for simultaneous measurement of a range of points on a target surface". U.S. Patent 5,075,561, 1991.

[66] M. Rioux and F. Blais, "Compact three-dimensional camera for robotic applications", *J. Optical Society of America A* 3:1518-1521, 1986.

[67] G. Roth and M.D. Levine, "Extracting geometric primitives", *CVGIP: Image Understanding* 58:1–22, 1993.

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

[69] Y.A. Shreider, *The Monte Carlo Method.* Pergamon Press, 1966.

[70] M.V. Srinivisan, S.B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of lateral inhibition in the retina", *Proc. Royal Society London B* 216:427–459, 1982.

[71] M.C. Steckner and D.J. Drost, "Fast cepstrum analysis using the Hartley Transform", *IEEE Trans. Acoustics, Speech, and Signal Processing* 37(8):1300–1302, 1989.

[72] C.V. Stewart, "A new robust operator for computer vision: Application to range data", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 167–173, 1994.

[73] C.V. Stewart, "A new robust operator for computer vision: Theoretical analysis", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 1994.

[74] M. Subbarao and N. Gurumoorthy, "Depth recovery from blurred edges", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 498–503, 1988.

[75] M. Subbarao and T-C. Wei, "Depth from defocus and rapid autofocusing: A practical approach", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 773–776, 1992.

[76] G. Surya and M. Subbarao, "Depth from defocus by changing camera aperture: A spatial domain approach", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 61–67, 1993.

[77] W. Teoh and X.D. Zhang, "An inexpensive stereoscopic vision system for robots", in *Proc. IEEE International Conference on Robotics*, pp. 186-189, 1984.

[78] D. Terzopoulos, "Multilevel computational processes for visual surface reconstruction", *Computer Vision, Graphics and Image Processing* 24:52-96, 1983.

[79] D. Terzopoulos, "The computation of visible-surface representations", *IEEE Trans. Pattern Analysis and Machine Intelligence* 10(4):417-438, 1988.

[80] S. Ullman, "Analysis of visual motion by biological and computer systems", *IEEE Computer* 14(8):57-69, 1981.

[81] Y. Yeshurun and E.L. Schwartz, "Cepstral filtering on a columnar image architecture: a fast algorithm for binocular stereo segmentation", *IEEE Trans. Pattern Analysis and Machine Intelligence* 11(7):759-767, 1989.