Bayesian Estimation of Stereo Disparity from Phase-Based Measurements

by

Jianping Wu

A thesis submitted to the
Department of Computing and Information Science
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
February 2000

Abstract

Phase-based methods have been successfully used in many areas of computer vision, such as stereo matching and optical flow estimation. Because of the many desirable properties of phase, phase-based methods have some advantages over other methods. Phase-based methods provide measurements of binocular disparity at a set of scales and spatial orientations, and at a set of pre-shifts. However, current computer vision techniques combine these estimates in a somewhat ad hoc way, assuming that left and right images are simple translations of one another. Phase-based binocular measurements are also thought to comprise the first stage of disparity processing in the primary visual cortex. However, the subsequent stages that combine the measurements to find a unique disparity map are unknown.

The goal of this thesis is to formulate the estimation of binocular disparity from a collection of phase-based measurements. Using a Bayesian probabilistic approach, the goal is to compute a probability distribution over disparities given a set of phase-based measurements. The main contributions of this thesis concern the development of the likelihood function. Additional contributions concern the combination of the likelihood functions from different channels. We investigated two prior models for this purpose. The first approach assumes an uninformative prior of the disparity field. The resulting algorithm combines the measurements by taking the product of the likelihood functions over scales and orientations. The second approach uses a multi-scale Markov model, which takes into account the spatial coherence of typical disparity fields. The latter approach allows us to avoid both the problems associated with coarse-to-fine methods, and the iterative nature of existing MRF based approaches.

Acknowledgements

I would like to express my gratitude my supervisors, Dr. David Fleet and Dr. Dorothea Blostein. They made my graduate study an invaluable experience that I will long remember. David's insights, wisdom, mixed with large dose of patience has made an otherwise tough topic into an exciting learning journey. Dorothea's perspective, generic interest in research, as well as her encouragement kept my thesis thesis a pleasant experience.

I am very grateful to Sean Cai, Bin Jin and Hong Wu for their friendship and support during my stay in Kingston. I also thank Yang Dong for her proof reading of my thesis. My thanks to Richard Zanibbi and Edward Lank for turning an otherwise garage like room into one of the best labs in the department. Special thanks go to Debby Robertson - she is simply indispensable to us graduate students. I also thank Tom Brashaw and Gary Powley for their technical support for the workstations and the network.

Finally, I would like to thank this department and Queen's university for their financial support.

Contents

Chapter 1

Introduction

One important function of the human vision system is to construct a representation of a three-dimensional scene from the two two-dimensional images obtained by the left and right eyes. Although with other monocular cues and prior knowledge, one can still tell the relative distance of objects with one eye closed, having two images allows one to infer depth information more accurately for many tasks. For example, putting a pen into its cap is a trivial task with two eyes open. However one may find it challenging with one eye closed. This ability of finding the depth information encoded within multiple images is called *stereo vision.* In particular, it is called *binocular stereopsis* when dealing with a pair of images.

Researchers have been interested in computationally implementing stereo vision for the past thirty years. Computational stereo vision has many potential applications, such as robot navigation, the control of autonomous land vehicles and the automation of factory assembly lines. Unfortunately, like most problems in computer vision, the stereo problem has proven to be more difficult than originally anticipated.

In remaining sections of this chapter, we introduce the principles underlying the recovery of stereoscopic depth, the complications associated with it, and some existing techniques that are commonly used. We also outline the main contributions of this thesis, namely, the development of a Bayesian phase-based method for estimating binocular disparity.

1.1 Theoretical Basis

Binocular stereopsis uses a pair of images taken by cameras from different viewpoints. In the special but most common case, the two images are taken by two parallel cameras. These cameras have

- the same vertical position, the same orientation, and
- identical focal lengths.

In what follows, we assume that the images are taken by such a camera pair. Assume that a point in the 3-D world (we call it a *scene point* hereafter) is projected to both image planes of the camera pair, [1] and we know the exact location of the image point on the image planes. Given the positions of the center of projection of the cameras, as illustrated in Fig. 1.1, we can determine the location of the scene point. This technique, called *triangulation*, forms two rays that go through the respective image points and centers of projections. The intersection is the location of the scene point.

The primary task of a stereopsis algorithm is to establish point correspondences. That is, for each point in one image, find a point in the other image, such that they are the projections of the same scene point. Solving the correspondence problem is

[1] A scene point may not be projected to both image planes if there exists occlusion. We discuss occlusion later in this chapter.

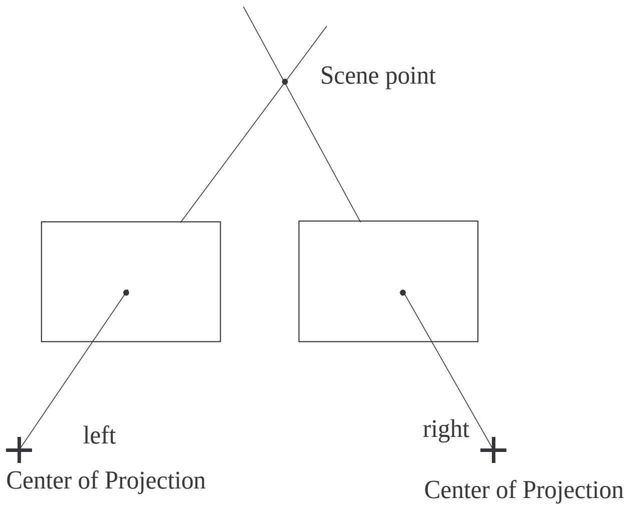

Figure 1.1: Triangulation. Given two image points that are the projection of the same scene point on different image planes, we can determine the three-dimensional location of the scene point. (After [Nal93].)

sometimes called *stereo matching*, because it involves matching points in two different images. It might appear that the establishment of correspondence requires that an entire image be searched for every point in the other image. Fortunately, such a two-dimensional search is unnecessary due to a powerful constraint: the *epipolar constraint* [KWS75]. As illustrated in Fig. 1.2, given an image point A, its corresponding point B in the other image is constrained to lie on the *epipolar line*, which is the projection of the straight line that goes through the scene point and point A.

In parallel camera configurations, the epipolar lines coincide with the horizontal scan lines of the images and are parallel [Nal93]. For a scene point A, as shown in Fig. 1.3, assume that its corresponding image point m in one image is at (u, v), and its corresponding image point m' in the other image is at (u', v'). The projection points m and m' have the same vertical position such that $v = v'$. **Binocular disparity** is a

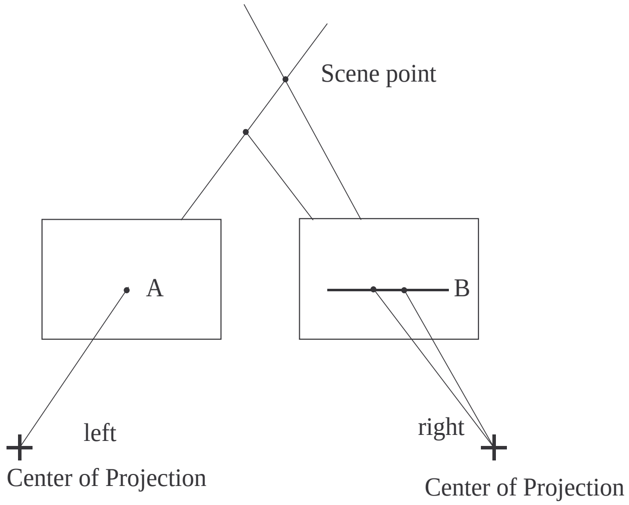

Figure 1.2: The epipolar constraint. (After [Nal93].)

term encountered frequently in stereopsis. In the parallel camera setting, the disparity denotes the displacement between the two corresponding image points, whose value d is defined as

$$d = u' - u \tag{1.1}$$

Given the camera calibration, we can derive the *depth measure* z of A from the disparity value d (Figure 1.4):

$$z = \frac{ft}{d} \tag{1.2}$$

where f is the camera focal length, and t is the distance between the two cameras. Since disparity encodes important depth information, it becomes the primary task of stereo vision to find the disparity map of the image pair. In this thesis, when we talk about stereo vision, we are referring mainly to the estimation of the disparity map.

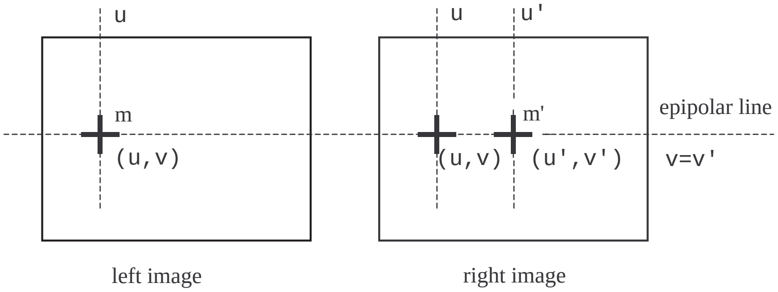

Figure 1.3: Epipolar constraint in parallel camera configurations. m' in the right image corresponds to m in the left image. m and m' are on the epipolar line and have the same vertical position.

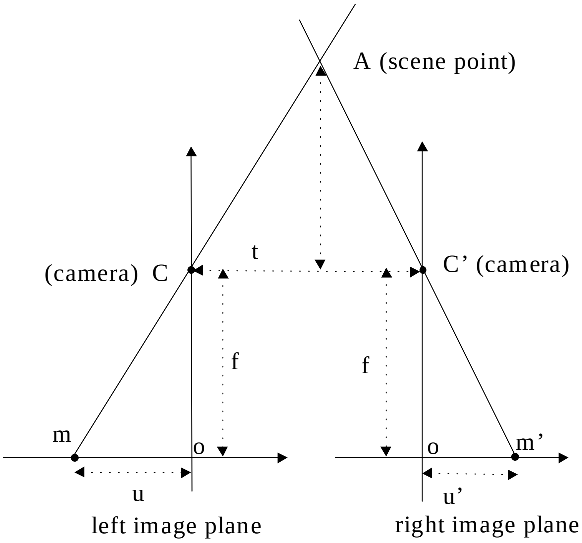

Figure 1.4: Given the camera calibration and disparity value, we can derive the depth measure z of the scene point A. t is the distance between the two cameras. f is the focal length.

In the general case, the cameras may not be in such ideal positions, but one can use the camera model to map the two images into a common rectification plane in which the epipolar lines are horizontal.

1.2 Correspondence Establishment

The task of matching points between the two images is known as the *correspondence* problem. This seemingly easy task is complicated by several factors:

- *Perspective Projection*: When projected to the image planes, a surface in a 3-D scene may undergo geometric deformations between left image and right images.
- *Uniform brightness*: Many images contain large regions of uniform brightness. The matching is often ambiguous for these regions.
- *Discontinuity and occlusion*: Many stereo scenes contain discontinuities in depth at object boundaries, discontinuities in surface orientation, and steeply sloping surfaces. Discontinuities often cause *half occlusion*, that is, some points in one image may not have corresponding points in the other image. This is illustrated in Fig. 1.5.
- *Noise*: The sources of noise include the quantization error, noise from the cameras, lighting variation between images, specular reflection, etc. Because of the noise, the feature (such as intensity, edge, and zero-crossing) values for corresponding points in the left and right images often differ.

Because of the different sources of intensity variation, most scene points will have some difference in image intensity between the two images. And because of occlusion,

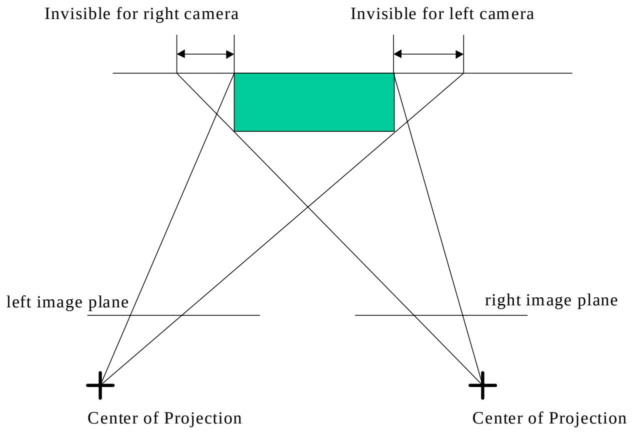

Figure 1.5: Discontinuity at object boundaries often causes *half occlusion*, that is, some scene points are visible from one viewpoint while invisible from the other viewpoint. (After [Nal93].)

there may be regions of *half occluded* points that appear in only one image and consequently have no match at all. For years, people have offered many solutions to the correspondence problem. However there are no satisfactory algorithms that adequately address all of the complications. In what follows, we introduce some existing techniques and compare their advantages and disadvantages.

The existing techniques can be classified as *Phase-Based*, *Intensity-Based*, and *Feature-Based*. We begin with a brief introduction of these methods.

1.2.1 Intensity-Based Methods

A straightforward approach to establishing correspondence along epipolar lines is to match points on the basis of their image intensities. It assumes that scene points have the same intensity in each image. However as we mention above, the image intensity corresponding to a 3-D point may not remain the same in the two images.

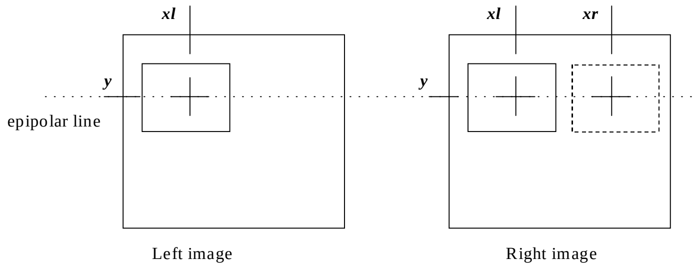

Figure 1.6: Window-based method. The window in the right image slides along the conjugate epipolar line to find the best match.

In addition, as several points in each image along the epipolar line may have the same intensity, establishing correspondence by matching intensities on a point-by-point basis may not be feasible, unless there is some form of smoothness constraint that limits the solution space [GLY95]. So instead of matching point-by-point, it has been common to use a *window-based* method as illustrated in Fig. 1.6. This method assumes that the disparities are, ideally, constant within a small window. As shown in Fig. 1.6, consider a small window centered at a point in the left image. Then along the eipolar line at the same vertical position in the right image, we slide a window of the same size and find the location where the two windows have the best match. The displacement between the left and right windows is considered the disparity at the window centers. The matching process uses a simple correlation scheme to measure the quality of match; that is, the more the two regions in the windows are correlated, the more likely they ought to be matched.

A problem associated with this window-based approach is that the size of the correlation windows must be carefully chosen, because it assumes constant disparity within the window. If the size is too small (the extreme case is that the window is one pixel), it may not capture enough image structure, and thus may be too sensitive to noise and have many false matches. A large window (the extreme case is that the window has the same size as the image) may result in a disparity map with a loss of fine detail because, for a large window, the assumption of constant disparity is often violated. In an *adaptive matching window* approach, Kanade and Okutomi [KO93] proposed that the size and shape of the matching window be chosen adaptively on the basis of a local evaluation of the variation in both the intensity and the previously estimated disparity. The idea of an adaptive window is that one can smooth the noise without smoothing over sharp variations in disparity.

Another approach is to use a *coarse-to-fine multi-scale* (*multi-resolution*) matching scheme. One can apply this approach to the intensity-based methods as well as phase-based and feature-based methods. With this approach, an initial guess of disparity is provided from a coarser scale. Then the images at the next finer scale are pre-shifted (warped) by the initial guess of disparities so that they are in rough alignment and with smaller disparities. This approach not only reduces ambiguous matching, but also reduces computation. However, this approach has a fundamental weakness. A poor estimate at the coarse scale leads to incorrect estimate at the fine scale, from which the algorithm cannot recover. A more detailed description of this approach can be found in Section 2.2 and elsewhere [MP77, Nis84, Bar89].

1.2.2 Feature-Based Methods

In the feature-based approach, the image pair is first preprocessed by an operator to extract *features* that are stable under the change of viewpoint. The matching process is then applied to the attributes associated with the detected features. The features one can use may be edges, line segments, curve segments, etc. One of the most important and widely used features is the edge. There exist many operators for finding edges in an image. For example, one can use the $\nabla^2 G$ operator followed by a detection of zero-crossings [Cas96]. The edge-based method is not useful in image regions without edges, and cannot generally get a dense disparity map since edges are often sparse in an image. This is also true for other features such as corners and line segments. Hence, edge-based methods or other feature-based methods are often used in conjunction with intensity-based methods.

1.2.3 Phase-based Methods

Phase-based methods estimate disparity from the phase information in band-pass filtered versions of the binocular images. In particular, disparity is defined as the shift necessary to align the phase values of the two signals.

Phase has several desirable properties [FJJ91]. One advantage is that phase is amplitude invariant, and therefore, these techniques are robust even when there exist lighting variations between the two images. Phase has also been shown to be robust when the left and right images are near affine deformations of one another [FJ93], which commonly exist when viewing 3D surfaces that are not frontoparallel.

One of the main advantages of phase techniques is that phase is predominantly linear [FJ93]. As a consequence the estimation of disparity can be reduced to the displacement of (nearly) linear functions. In this way, it is possible to obtain dense disparity maps with sub-pixel accuracy, without requiring explicit sub-pixel signal reconstruction. However, phase is only uniquely defined over one wavelength of a band-pass signal, and therefore phase-based measurements from a single band-pass filter (channel) can only uniquely determine disparities up to half a wavelength. Thus, in practice, it is common to pre-shift the images using a set of plausible *pre-shifts* or using an initial disparity estimate from another channel at a coarser scale where wavelengths are longer, and hence they can measure larger disaprities, at the cost of poorer spatial resolution.

Finally, it is also interesting to note that neurophysiological research has shown that the first stages of disparity processing in the primary visual cortex in cats, primates, and in the visual wulst in owls are thought to use a phase-based measurement [DOF91, WF93]. This model is often referred to an energy model. In short, the energy model involves the cross-correlation of band-pass signals from the two eyes, from different (shifted) retinal positions, much like the phase-based disparity measurements of [JJ94, Fle94].

1.3 Thesis and Contributions

Phase-based methods provide measurements of binocular disparity at a set of scales and spatial orientations, and at a set of pre-shifts. Current computer vision techniques combine these estimates in a somewhat ad hoc way, assuming that left and right images are simple translations of one another. Phase-based binocular measurements are also thought to comprise the first stage of disparity processing in the primary visual cortex. However, the subsequent stages that combine the measurements to find a unique disparity map are unknown.

The goal of this thesis is to formulate the estimation of binocular disparity from a collection of phase-based measurements. Using a Bayesian probabilistic approach, the goal is to compute a probability distribution over disparities given a set of phase-based measurements. Using Bayes' rule, this posterior distribution can be expressed as the product of a likelihood function and a prior density function. The likelihood function specifies the probabilistic relation between the phase-based measurements and the underlying disparity field, and the prior model specifies our prior belief in the structure of disparity fields.

The main contributions concern the development of the likelihood function. Additional contributions concern the use of a multi-scale prior model. This resulting approach allows us both to avoid the problems associated with coarse-to-fine methods, and the iterative nature of existing Markov Random Field (MRF) based approaches.

Contributions related to the likelihood function include:

- Identification and modeling of the sources of variability in phase-based measurements as they arise from filter outputs. (Section 4.1 and Section 4.2)
- Empirical derivation of the form of the likelihood function for single phase-based measurements at multiple scales/orientations. We fit the form of the likelihood with a parameter model invariant in scale. (Section 4.3 and Section 4.4)
- Formulation of a joint likelihood function for a measurements at different preshifts for a single scale/orientation. (Section 4.5)

Initial step towards multi-scale combination:

- Implementation of the [FWH96, Fle94] method of summing binocular phase measurements over channels. (Section 2.5.2)
- Implementation of the simplest way to combine the measurement over scales and orientations, taking the product of the joint likelihood function across scales and

orientations. By doing so, we assume the independence of the measurements at different scales and orientations, and the uniform prior over the disparity map. (Section 5.1)

- Development of an algorithm based on a multi-scale prior model that prefers smooth disparity fields and small disparities. Implementation and testing of it against other algorithms. The new algorithm has many potential advantages over existing MRF-based approaches. Currently, many MRF-based algorithms incorporate smoothness models that require iterative procedures, with coarse-to-fine propagation of estimates. They are usually slow to converge and are therefore not suitable for real-time applications. The multi-scale algorithm here does not assume coarse-to-fine and it is computed in fixed time in terms of the number of pixels. (Section 5.4)

Our main goal for the multi-scale combination is to show the feasibility of using phase-based measurements in a Bayesian approach. Therefore, it is not our major concern to achieve a significant improvement over existing stereo matching methods at this early stage, nor do we provide in-depth comparison of our results with those from other existing methods. Nevertheless, the methods show potential for further improvement.

This thesis is organized as follows. Chapter 2 describes the phase-based technique on which our algorithm is based. Chapter 3 introduces the Bayesian approach in computer vision, as well as the likelihood function and prior model used in this approach. It especially focuses on the Markov Random Field (MRF) as a prior model and dicusses the choice of prior models. In Chapter 4, we explicitly model the sources of variability of the phase-based measurements. Using these models of variability, we empirically determine the likelihood function with a parameter model. Chapter 5 shows how to combine the likelihood functions obtained at different scales to reach an optimal estimate of disparity.

Chapter 2

Phase-Based Methods

This chapter reviews several important phase-based methods, namely, phase-difference, phase-correlation and the local weighted phase-correlation method. Phase-based methods have found application in image matching, including stereo matching and optical flow estimation. Despite the existence of numerous techniques that are phase-based, they all share a common feature, that is, they exploit the phase behavior in band-pass filtered versions of different views of a 3-D scene [FJ90, FJJ91, San88, Wen94, JJ94, Fle94].

Researchers have discovered many desirable properties of these techniques. The obvious advantage of phase-based methods over intensity-based ones is that phase is amplitude invariant, hence the measurement is robust even if there exist lighting variations resulting from, for example, surfaces with specular reflection. Another reason to use phase is that phase is predominantly linear [FJ93] in space. The phase linearity is important since it is easy to estimate the displacement of linear functions. In terms of the quality of disparity map, in phase-based methods, matching can exploit all phase values such that a dense set of estimates can be extracted by making full use of the available signal. Disparity estimates are obtained with sub-pixel accuracy, without requiring explicit sub-pixel signal reconstruction or sub-pixel feature detection and localization.

One of the most important advantages of phase-based approaches is the stability of band-pass phase behavior with respect to image deformations that typically exist between the left and right image [FJ93]. The stability of phase behavior means that a small affine deformation of the image causes a similar deformation of the phase. Although phase deformations do not exactly match input deformations, they are usually close enough to provide reasonable measurements for stereo matching and other vision applications.

One might be tempted to place more importance upon the amplitude than phase, since amplitude exhibits some recognizable structure. However, while the amplitude specifies the magnitude and significance of the band-pass filter output, it is the phase information that specifies the local structure of the response. This property of phase is particularly important in addressing the stereo matching and optical flow estimation. Neurophysiological data also imply the importance of phase information. For example, Ohzawa *et al.* [ODF90, FWH96] suggested that disparity sensitivity of neurons in the visual cortex might be a result of interocular phase shifts.

Central to phase-based methods are the filters that decompose the images into band-pass signals. In this chapter, we will first discuss the quadrature-pair filters that can be used to extract phase. Then we review the phase-difference and phase-correlation methods. After outlining the advantages and disadvantages of these two methods, we then introduce a method proposed by Fleet in [Fle94] that combines desirable properties of both methods.

2.1 Quadrature Filters

A pair of filters is said to be in quadrature phase if they have the same amplitude spectra, but differ in phase by 90°, that is, they are Hilbert transforms of each other. The Hilbert transform is given by

$$g(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(y)}{y - x} dy \tag{2.1}$$

As an example, Fig. 2.1A shows one filter of a quadrature filter pair, whose impulse response is second derivative of a Gaussian(G2), Fig. 2.1B is an approximation to the Hilbert transform of G2, which is usually called H2 [FA91]. G2 and H2 have the same frequency response as shown is Fig. 2.1C, but their phase responses differ by 90°.

In practice, we may use band-pass filters tuned to different orientations and scales to compute the disparity. One way to design and implement the oriented filters is to use the *steerable filter* [FA91]. See appendix A for details.

2.2 Phase-Difference Methods

Existing phased-based methods can generally be classified into two categories: phase-difference [JJ94, FJJ91] and phase-correlation methods [KH75]. In phase-based methods, disparity is defined as the shift necessary to align the phase values of band-pass filtered versions of two signals.

Let $I_l(x)$ be the left input signal and $I_r(x)$ be the right input signal.[1] Assume that they are filtered by complex-valued filters, such as a quadrature pair filter. Let the complex-valued responses of a single band-pass filter be $L(x)$ for the left signal

[1] Here the left and right signals are in 1-D because we assume they are on the epipolar lines.

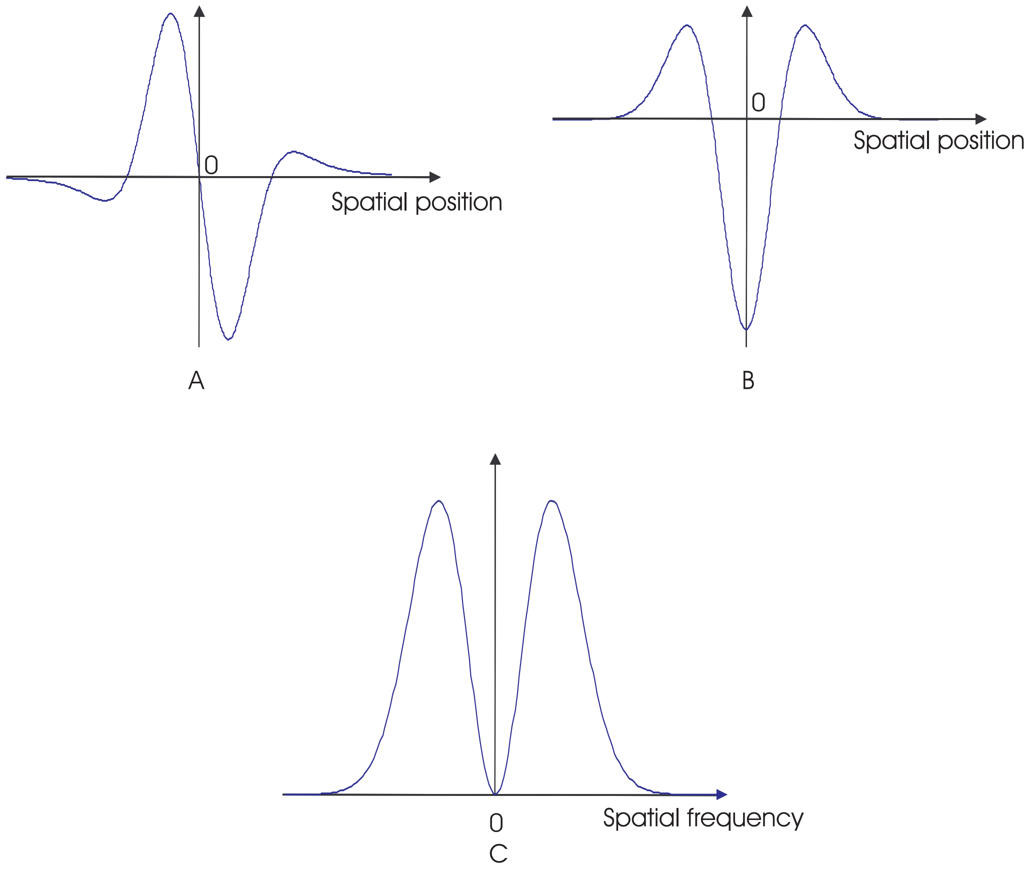

Figure 2.1: G2 and H2 filters and their frequency and phase responses. Fig. A shows a G2 filter, whose impulse response is second derivative of Gaussian, Fig. B is the Hilbert transform of G2, which is usually called H2. G2 and H2 have the same frequency response shown is Fig. C.

and $R(x)$ for the right signal. In what follows, we express L and R in terms of local amplitude and phase:

$$L(x) = A_l(x)e^{i\Phi_l(x)}, R(x) = A_r(x)e^{i\Phi_r(x)} \tag{2.2}$$

where $A_l(x)$ and $A_r(x)$ are amplitudes and $\Phi_l(x)$ and $\Phi_r(x)$ are phases. $L(x)$ and $R(x)$ could be outputs of 2-D filters, but here they are treated as a function of position x along the epipolar line.

The local image disparity, at a specific location x, is defined to be the shift $d(x)$ such that

$$\Phi_l(x - \frac{d(x)}{2}) = \Phi_r(x + \frac{d(x)}{2}) \tag{2.3}$$

Because phase is only uniquely defined in the interval $(-\pi, \pi]$, this approach can only measure disparities up to half a wavelength. If the disparity is too large then the computed phase difference can be wrong by a multiple of 2π. From filtered outputs with short wavelengths we can only measure small disparities. With large wavelengths we can measure large disparities; however we are unable to detect fine details in $d(x)$, because large wavelength means large filter support, which causes the loss of spatial resolution.

In order to use filters tuned to higher frequencies, so that we can achieve a more detailed disparity map, we need to use coarse and fine estimates. For example, it is common to use some form of coarse-to-fine control strategy [MP77, JJ89]. An initial guess of disparity is provided from a coarse scale where the filters output has a larger wavelength (the wavelength should be more than twice the largest expected disparity). Then the images at the next finer scale are pre-shifted (warped) by the initial guess of disparities so that they are in rough alignment and with smaller disparities. We can then use filters tuned to higher frequencies to determine how much further alignment is needed to bring them into exact match.

Assume that at a certain scale, the initial guess is sufficiently good and the filter outputs have frequencies close to the frequency ω_0 to which the filter is tuned to. Then the shift required to bring the left and right signals into match is given by

$$\widetilde{d}_0(x) \equiv \frac{[\Phi_l(x) - \Phi_r(x)]_{2\pi}}{\omega_0} \tag{2.4}$$

where $[\Psi]_{2\pi}$ denotes the principal part of Ψ that lies between $-\pi$ and π. However, if the outputs are not purely sinusoidal, then the disparity estimates $\widetilde{d}_0(x)$ will not be exact since the phase is not purely linear and the frequency is not constant.

In [FJJ91], Fleet suggests that we adopt a more general model. Instead of using the tuning frequency of the filter output, we should use the local frequency of the band-pass signal at each spatial location, i.e. the *instantaneous frequency.* The instantaneous frequency is defined as

$$\bar{\omega}_l = \frac{d\Phi_l(x)}{dx}, \bar{\omega}_r = \frac{d\Phi_r(x)}{dx} \tag{2.5}$$

Replacing ω_0 in Eq. (2.4) by the average instantaneous frequency between the left and right signals, we obtain an new estimate of the shift

$$\widetilde{d}_1(x) \equiv \frac{[\Phi_l(x) - \Phi_r(x)]_{2\pi}}{\overline{\omega}} \tag{2.6}$$

where $\overline{\omega} = \frac{1}{2}(\bar{\omega}_l + \bar{\omega}_r)$.

[FJJ91] gives an estimate of the disparity error of Eq. (2.6). Assume that left and right filter outputs are shifted versions of one another with disparity $d(x) = \delta$, then the phase difference in Eq. 2.6 is

$$\Delta\Phi(x) = \Phi_l(x) - \Phi_r(x) = \Phi(x + \frac{\delta}{2}) - \Phi(x - \frac{\delta}{2}) \tag{2.7}$$

If $\Phi(x)$ is smooth, we can rewrite $\Phi(x+\frac{\delta}{2})$ and $\Phi(x-\frac{\delta}{2})$ as Taylor series about x:

$$\begin{aligned}\Phi(x+\frac{\delta}{2}) &= \Phi(x)+\frac{\delta}{2}\Phi^{'}(x)+\frac{\Phi^{''}(x)}{2}(\frac{\delta}{2})^2+O(\delta^3\Phi^{'''}(x))\\ \Phi(x-\frac{\delta}{2}) &= \Phi(x)-\frac{\delta}{2}\Phi^{'}(x)+\frac{\Phi^{''}(x)}{2}(\frac{\delta}{2})^2+O(\delta^3\Phi^{'''}(x))\end{aligned} \tag{2.8}$$

So $\Delta\Phi(x)=\delta\Phi^{'}(x)+O(\delta^3\Phi^{'''}(x))$. Then the disparity error $\epsilon(x)$ for the new estimate $\widetilde{d}_1(x)$ is given by

$$\epsilon(x)=O\Big(\frac{\delta^3\Phi^{'''}(x)}{\Phi^{'}(x)}\Big) \tag{2.9}$$

2.3 Measurement of Phase Differences

There are several ways to measure phase differences. For example, one can take the complex-valued product of left output and the complex conjugate of the right [JJ89]:

$$C(x)=L(x)R^*(x)=A_l(x)A_r(x)[\cos\Delta\Phi(x)+i\sin\Delta\Phi(x)] \tag{2.10}$$

where $\Delta\Phi(x)=\Phi_l(x)-\Phi_r(x)$. The real and imaginary part of $C(x)$ can be computed directly from the real-valued filter outputs:

$$A_lA_r\cos\Delta\Phi = \mathrm{R}e[L]\mathrm{R}e[R]+\mathrm{I}m[L]\mathrm{I}m[R] \tag{2.11}$$

$$A_lA_r\sin\Delta\Phi = \mathrm{I}m[L]\mathrm{R}e[R]-\mathrm{R}e[L]\mathrm{I}m[R] \tag{2.12}$$

In the discussion so far, we assumed the phase $\Phi(x)$ is smooth and stable. However, the phase signals are occasionally very sensitive to spatial position and variation in scale [FJ93]. This instability occurs in the neighborhoods of phase singularities, where the amplitude of the signal goes through the origin in the complex plane. It is necessary that the singularity neighborhoods be detected so that incorrect disparity estimate can be avoided. This can be done with constraints on the instantaneous frequency and the amplitude derivative of the filter output [FJ93].

2.4 Global Phase-Correlation

Phase-correlation methods use Fourier phase for signal registration. They assume that in a *small local area*, one image is simply a shifted version of the other image.

$$I_r(x) = I_l(x - d) \tag{2.13}$$

By using the Fourier shift theorem, taking the Fourier transform of both sides of Eq. (2.13), yields $\hat{I}_r(\omega) = \hat{I}_l(\omega)e^{-i\omega d}$, where $\hat{I}_l(\omega) = A_l(\omega)e^{i\Psi_l(\omega)}$, and $\hat{I}_r(\omega) = A_r(\omega)e^{i\Psi_r(\omega)}$. $A_l(\omega)$ and $A_r(\omega)$ are the amplitude spectra, while $\Psi_l(\omega)$ and $\Psi_r(\omega)$ are the phase spectra. $\omega d = \Psi_l(\omega) - \Psi_r(\omega)$, which represents the difference in the phases of the respective Fourier coefficients at each frequency ω.

Taking the product of the left Fourier spectra and the complex conjugate of the right, and then dividing by the product of their amplitude spectra, we obtain

$$\frac{\hat{I}_l(\omega)\hat{I}_r^*(\omega)}{A_l(\omega)A_r(\omega)} = \frac{A_l(\omega)A_r(\omega)e^{i(\Psi_l(\omega)-\Psi_r(\omega))}}{A_l(\omega)A_r(\omega)} = e^{i\omega d} \tag{2.14}$$

where $\hat{I}_r^*(\omega)$ is the complex conjugate of $\hat{I}_r(\omega)$. The inverse Fourier transform of $e^{i\omega d}$ is $\delta(x + d)$, where $\delta(\cdot)$ is the Dirac delta function. So the phase-correlation methods measure disparity by finding peaks in

$$F^{-1}\left[\frac{\hat{I}_l(\omega)\hat{I}_r^*(\omega)}{A_l(\omega)A_r(\omega)}\right] \tag{2.15}$$

In practice, the disparity is measured locally, that is, by using the windowed regions of the left and right images instead of the whole original images. The size of the window must be larger than the expected displacement so that there is sufficient information in two windows that can be used for matching.

The phase-correlation method determines the disparity based on the consistency of information at different scales and orientations. It does not require a coarse-to-fine control strategy and it can work even when the band-pass signals are shifted by more than half a wavelength of the lowest frequencies in the signal.

Fleet showed [Fle94] that one can view the phase-correlation methods as using a voting scheme to find the disparity. The inverse Fourier transform is the reconstruction of a function by summing up all the sinusoids weighted by their amplitudes. So the inverse Fourier transform is a sum of phase-shifted sinusoidal functions. Ideally, there will be a single disparity at which peaks coincide across a wide range of frequencies to form a unique peak.

2.5 Local Weighted Phase-Correlation

So far, we have discussed the phase-difference method and the phase correlation method. Both methods have their advantages and disadvantages. Phase-difference methods have many desirable properties. In recent years, the use of local wavelet filters and the stability constraints greatly improve the robustness of the measurement. However, the phase-difference method usually requires some form of coarse-to-fine control strategy, which is often regarded as unsatisfactory. If a poor estimate is obtained at the coarsest scale, the next finer scale will have a poor initial guess, which will bring the two images into false registration and the rest of the process may converge to the incorrect disparity. In addition, there is growing evidence to indicate that the correspondence search in human stereo vision may not be a coarse-to-fine process [MDA94].

The major advantage of the phase-correlation method is that the voting scheme it uses to determine the disparity is based on the consistency of information at different scales and orientations. It does not require a coarse-to-fine control strategy and it works even when the band-pass signals are shifted by more than half the wavelength of the lowest frequency. Some researchers have found fundamental weaknesses in using the Fourier methods to measure the relative phase shift between the left and right images [KH75, Sto86]. The wrap around effect is inevitable since a window-based scheme is used here. Note that this problem cannot be easily solved by using different windowing function [Sto86].

Fleet [Fle94] proposed the Local Weighted Phase-Correlation method that combines the robustness of phase-difference methods and the voting scheme of phase-correlation methods. This method uses a measurement of local phase-difference proposed in [JJ94]. In this section, we first introduce this measurement and discuss some of its important properties. Then we describe how this measurement is utilized in the Local Weighted Phase-Correlation method.

2.5.1 The Measurement of Local Phase-difference

In 2.2, in order to compute the phase difference, we directly take the normalized product of left and right outputs as in Eq. (2.10). Here we introduce a different method proposed in [JJ94] to measure the local phase difference. Assume that the left and right signals have been brought into rough alignment by a pre-shift τ of one signal, for example, the right one. We need to know how much more alignment is needed to bring them into an exact match. To do so, we can find a complex scalar z to minimize the squared difference between the left and the shifted version of the right signal, i.e.,

$$\int W(x)|L(x) - zR(x+\tau)|^2 dx \tag{2.16}$$

where $W(x)$ is a small, localized Gaussian window. Motivated by this purpose, Jenkin and Jepson proposed a new measurement of local phase difference [JJ94]

$$C(x,\tau) = \frac{W(x) \otimes [L(x)R^*(x+\tau)]}{\sqrt{W(x) \otimes |L(x)|^2}\sqrt{W(x) \otimes |R(x)|^2}} \tag{2.17}$$

where $\otimes$ is the convolution operator. The introduction of the localized window provides more stable measurement, which we will discuss later. One can show that the phase of $C(x,\tau)$ corresponds to the phase of z, which minimizes Eq. (2.16). There are several important properties of this measurement.

First, the phase of $C(x,\tau)$, as the phase of $C(x)$ in Eq. (2.10) is a phase difference that encodes the shift required to match the phases of the left and right band-pass signals. In addition, the peaks in the real part of $C(x,\tau)$ can be used as votes for candidate disparities τ between left and right filter outputs at location x.

The second important property of $C(x,\tau)$ is that its magnitude provides a confidence measure for the goodness of fit between the phase-shifted left and right signals. At some location x_0, $C(x_0,\tau)$ can be rewritten as the local spatial average of vectors inside the localized window (ignoring the window weights for convenience)

$$C(x_0,\tau_0) \approx \frac{\sum A_l A_r e^{j\Delta\Phi}}{\sqrt{\sum A_l^2}\sqrt{\sum A_r^2}} \tag{2.18}$$

where each vector in the localized window has magnitude $A_l A_r$ and phase $\Delta\Phi$. $\sum A_l A_r e^{j\Delta\Phi}$ will be large in magnitude when the phases among all vectors have no or little variation. If the phases, $\Delta\Phi$, among these vectors vary significantly (that is, the vectors are in the singularity regions), the vectors will cancel each other when they are summed up. In this way the magnitude of $C(x,\tau)$ depends on the local consistency of the phase difference within the window. On the other hand, when all phase differences are the same, the magnitude of Eq. (2.18) depends on the cross-correlation of the amplitudes of the left and right filter outputs. So the magnitude of $C(x_0, \tau)$ depends both on the local stability (consistency), and the cross-correlation of the amplitudes of the left and right filter outputs. One can see that the localized window serves as a means to alleviate the effect of phase instability to the singularity regions [FJ93].

The real part of $C(\tau, x)$ can be used to vote for candidate disparities between left and right filter outputs at some location x. We expect a peak value at the true disparity. However, the peaks may also occur elsewhere besides the true disparity. Fig. 2.2 shows the real part of $C(\tau, x)$ at some location x for a stereogram pair. The true disparity is known to be -1 pixel, but there are also peaks elsewhere. These peaks, which are called *false peaks*, occur because phase signals cycle between $-\pi$ and π as a function of spatial position [FWH96]. The peak occurs when the phase difference $\Delta\Phi(x)$ is zero. If the left phase signal at location x is $\Phi_l(x)$, then the phase at the location $x + \lambda$ is expected to be almost the same, where λ is the wavelength of the filter output. So the left phase signal will usually equal the right phase signal at several spatial locations, which causes the false peaks. The false peaks sometimes may be larger than the peaks at the true disparity (correct peaks). From Eq. (2.18), one can see that the value of the peak depends on the value of the left and right amplitudes. When the amplitudes are larger at the false peak than at the correct peak, the false peak is larger than the peak at true disparity. The existence of false peaks means the binocular measurement alone cannot be used as ideal disparity detector. In next section, we describe a multi-scale approach proposed by Fleet in [Fle94] to overcome the effect of the false peaks.

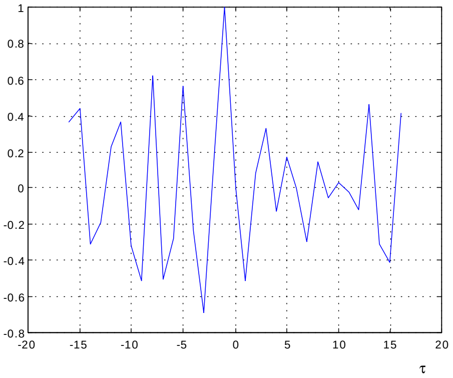

Figure 2.2: The real part of $C(\tau, x)$ at some location x for a pair a stereogram. The true disparity is known to be -1 pixel, but there are also peaks elsewhere. These peaks are called *false peaks*.

2.5.2 Disparity from Local Weighted Phased-Correlation

The local weighted phased-correlation method borrows its ideas from the phase-difference and phase-correlation methods. The Fourier transforms in phase-correlation are replaced with a set of quadrature-pair filters tuned in different orientation and scale with constant octave bandwidth. The filter outputs are used to compute the voting functions described above, with a series of pre-shifts. The resulting voting functions are then summed across different orientations and scales, from which the disparity measurements are extracted.

We have implemented this method as follows. First, we construct a three-scale Gaussian pyramid from the original images, sub-sampled at each level by a factor of 2 horizontally and vertically. Three quadrature-pair filters are applied at each scale, tuned to orientations 0^o, $+45^o$, and -45^o. Let $C_{s,j}(x, \tau)$ denote the binocular measurements obtained from the filter outputs of each scale and orientation using

Eq. (2.17), where subscript s refers to the sth scale, and j refers to the jth filter. The use of these measurements, as in[Fle94], involves a simple summation

$$S(x,\tau) = \sum_{j,s} C_{s,j}(x,\tau) \tag{2.19}$$

At each scale, the range of x is different by a factor of 2 both vertically and horizontally from its next finer scales. The range of pre-shift τ is also different by a factor of 2, since the disparity is linearly scalable over scales. To perform the summation, we have to interpolate in x to bring the measurements back to the resolution of the original image. We also need to interpolate $C_{s,j}(x,\tau)$ in τ for all the scales except the finest ones.

In [Fle94], Fleet shows that $C(x,\tau)$ is expected to be band-pass in τ and low-pass in x. Thus the interpolation in x is done by linear interpolation. The interpolation in τ cannot be done by linear interpolation since $C(x,\tau)$ is band-pass in τ. Otherwise, this may result in aliasing error and cause the missing of peaks. This is illustrated in Figure 2.3, where simple linear interpolation may cause the peak value to appear in an incorrect pixel position.

One way to overcome this problem is demodulating the band-pass signal so that it becomes low-pass before the interpolation. The centre frequency of $C_{s,j}(x,\tau)$ is close to the filter's tuning frequency. One can demodulate the signal by multiplying it with a sinusoidal signal with the filter's tuning frequency. This results in a low-pass signal. Then one can linearly interpolate the low-pass signal, followed by modulation to undo the initial demodulation. Figure 2.3 shows that the demodulation and modulation method improves the accuracy of measurement.

Given the $S(x,\tau)$, near the true disparity we expect to find a zero in its imaginary part and a peak in its real part. The summation of $C(x,\tau)$ over multiple scales can

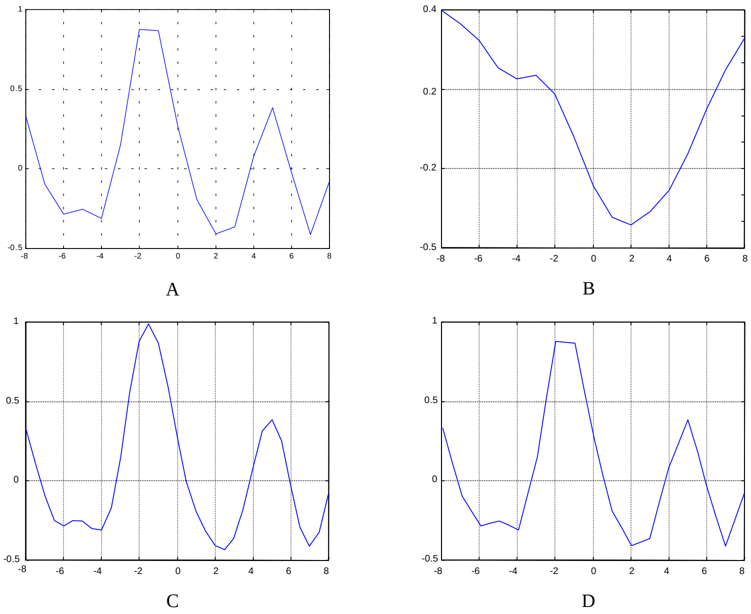

Figure 2.3: (A) Real part of a curve of binocular measurement at some position. (B) Modulate the curve in (A) and perform the interpolation. (C) Demodulate the curve in (B) and we obtain an interpolated version of curve in (A). (D) Perform the linear interpolation without demodulation and modulation. One may notice that it misses some peaks.

effectively eliminate the false peaks that occur on a single scale. An example taken from the random dot stereogram is shown in Fig. 2.4. One can see it also enhances the peak near the correct disparity. The expected interval between false peaks is approximately the wavelength of the filters applied on the scale. Thus the false peaks at different scale occur at different disparities. Summation over enough scales yields a prominent peak only at the true disparity.

Summation over orientations also helps to enhance the correct peak and attenuate false peaks. When input images contain textured elements, such as a textured surface or random dot stereograms, the chance of false peaks is high. In this case, the filters

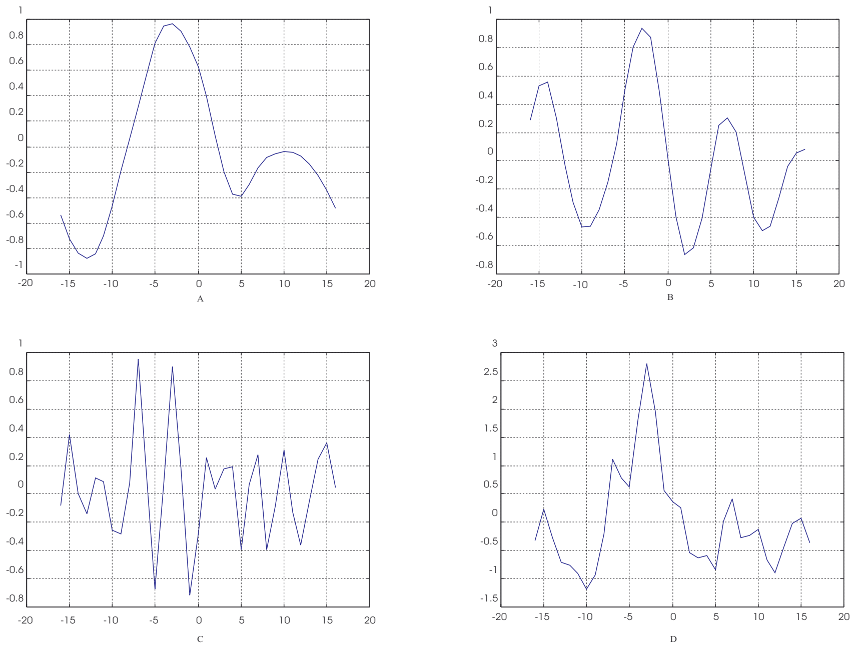

Figure 2.4: Each curve in (A), (B) and (C) is the real part of the binocular measure for a position x at a different scale, with (A) the coarsest scale and (C) the finest scale. Curve in (D) is the summation of curves across the scales. When peaks coincide across scales, the summation enhances the peak while at other disparities the peaks cancel each other.

tuned to different orientations provide nearly independent responses. Therefore, false peaks are expected to occur at different disparities and can be cancelled through summation.

The algorithm has been tested on several real image pairs and some of the results are shown here. Fig. 2.5 shows the standard image pair of the Pentagon building as seen from the air. Fig. 2.6 is the recovered disparity map with half pixel resolution[2]. Fig. 2.7 shows frame 2 and frame 4 of the SRI tree sequence. Fig. 2.8 shows the disparity estimates using these two frames with half pixel resolution. Fig. 2.9 is the disparity map with the sub-pixel resolution obtained using linear interpolation of the zero-crossing in the imaginary part of $S(x, \tau)$.

The method we discuss above is one simple way to utilize the binocular measurements from different scales and orientations and it achieves reasonably good result. It has also been used to model the neural encoding of binocular disparity [FWH96].

2.6 Discussion

The local weighted phase-correlation is an effective way to eliminate the false matching by summing of information over scales. However there are some problems with it. First, it gives every scale and orientation the same weight when performing the

[2] This means the binocular measurements are obtained every half pixel. One can shift an image by a value Δd less than one pixel through interpolation. Take linear interpolation as example, if the original image is $I(x, y)$, shifting the image by Δd results a new image $\hat{I}(x, y)$:

$$\hat{I}(x, y) = I(x+1, y)\Delta d + I(x, y)(1 - \Delta d)$$

One can use cubic or spline interpolation to achieve more accurate result. We use cubic interpolation in our experiment.

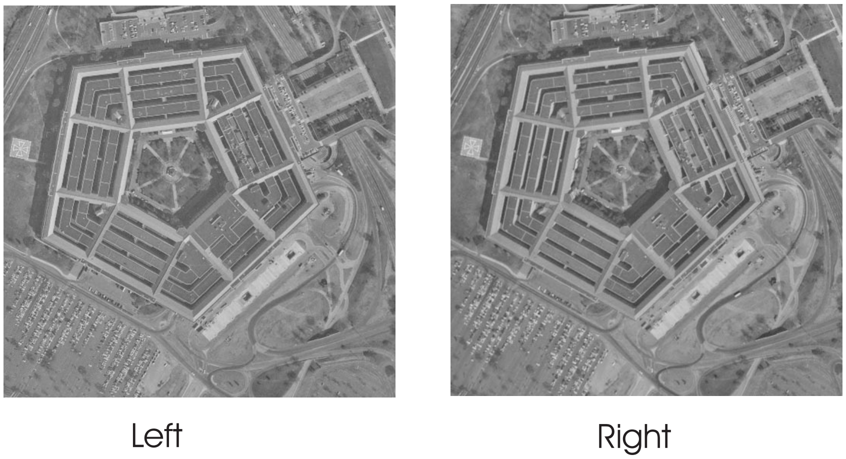

Figure 2.5: The standard image pair of the Pentagon building as seen from the air.

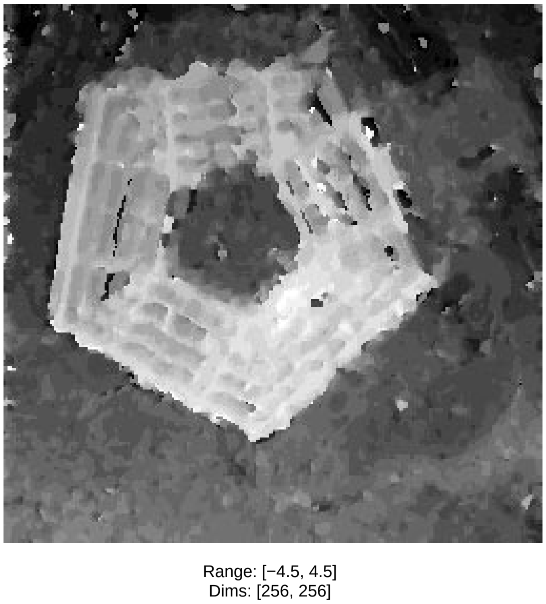

Figure 2.6: The recovered disparity map from the Pentagon image pair with half pixel resolution.

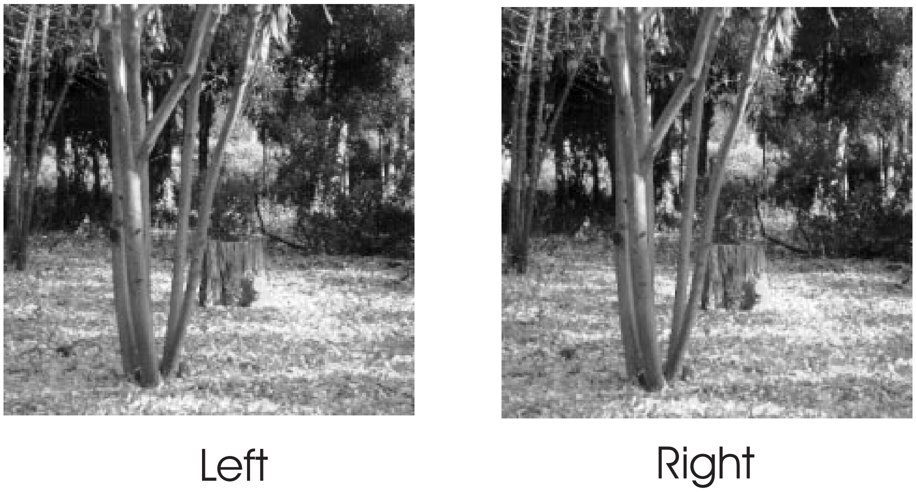

Figure 2.7: The SRI tree sequences: frame 2 and frame 4.

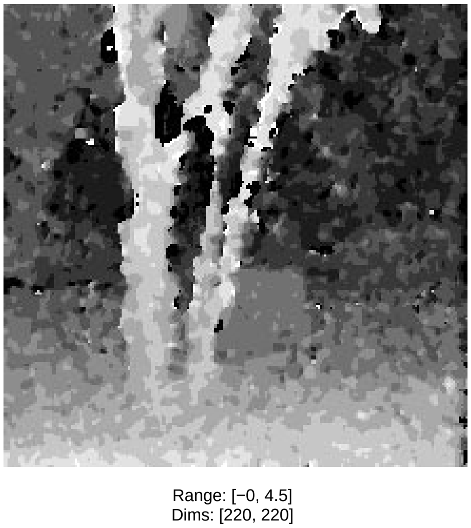

Figure 2.8: SRI tree sequences: Disparity estimates using the local weighted phase-correlated method with half pixel resolution.

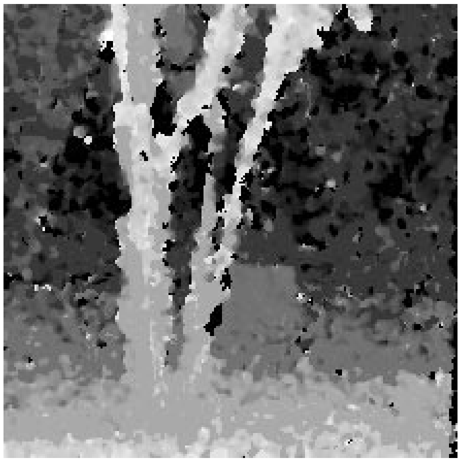

Figure 2.9: SRI tree sequences: Disparity estimates using the local weighted phase-correlated method. The sub-pixel resolution is obtained using linear interpolation of the zero-crossing in the imaginary part of $S(x, \tau)$.

summation. However, note that the coarse scales have to be interpolated before being added to the fine scales. The interpolation does not provide any new information, because the interpolated results are derived from known information. It may not be appropriate to give the values from the interpolation the same weight as the original values at the fine scale. In addition, there is no physiological evidence for summation over scale and orientation in the visual system to build disparity detectors. Nevertheless, this method indicates the promise of the multi-scale and multi-orientation approach. The problem is how to effectively use the measurements at different scales and orientations.

The other way of pooling information over scales is the coarse-to-fine approach. We have already discussed the weakness of this approach. A poor estimate at coarse scale will provide an incorrect initial guess for the estimate at fine scale, which can let the process converge to incorrect disparity. In this thesis, we investigate how to appropriately utilize information at different scales in the Bayesian approach.

2.7 Biological Vision

Besides the many advantages of phase-based methods for stereo matching over other methods, Neurophysiological research has also shown that phased-based measurements comprise the first stage of disparity processing in the primary visual cortex of many mammals and in the visual wulst of the owl [DOF91, WF93]. Therefore, phase-based methods are not only important for computer vision research, they also play important role in modeling the biological stereopsis. The binocular measurement introduced in this chapter is particularly interesting, since it has been used to model the response of V1 neurons to disparities [FWH96].

Despite the success of phase-based method in modeling the early stage of disparity processing, the subsequent stages that combine the measurements to find a unique disparity map are unknown. The commonly used coarse-to-fine control strategy in computer vision may not be suitable for modeling this process. That is because with the coarse-to-fine approach, a poor estimate at the coarse scale leads to incorrect estimate at the fine scale, from which the algorithm cannot recover. There is also evidence against the use of coarse-to-fine control strategy in biological vision [MDA94]. In [FWH96], the second stage of disparity processing is modeled as linear pooling, similar to the summation approach in the local weight phase-correlation method. However, there is little evidence for pooling over scale or orientation in the primary visual cortex. It is our hope that the Bayesian approach to information pooling, which is investigated in this thesis, may be helpful to understanding the second stage of disparity processing in biological visual system.

Chapter 3

Bayesian Approach

In this chapter, we introduce the Bayesian approach as a framework to solve computational vision problems, especially in the context of binocular stereopsis. In computer vision, the most often encountered question is how to infer the scene properties from the given image data. The image data is usually a 2-D representation of the scene. However the representation is often ambiguous, due to the fact that images are 2-D and often corrupted by noise. The Bayesian approach provides a formal framework to solve the ambiguity. One needs both a prior model and a likelihood function to apply the Bayesian approach. In this chapter, we briefly introduce some existing techniques of deriving the likelihood function. We also discuss the Markov Random Field (MRF), which is one of the most popular prior models used in computer vision.

3.1 The Bayesian Formulation of Vision Problem

The basic idea behind the Bayesian approach is to balance the information provided by the image data with the prior expectation of the scene. Let the image data be $\mathbf{I}$, and the scene properties of interest be $\mathbf{S}$ (e.g., object motion, surface shape, disparity etc.) In a Bayesian approach, we assume that scenes have some common statistical properties, which are represented by the prior probability distribution $P(\mathbf{S})$. $P(\mathbf{S})$ encodes our expectation about the scene. For example, if we expect that the surfaces of the scenes are generally smooth, then we would assign a low prior probability to a scene with steep slopes or creases.

We can think of the image data as a projection Π (*image formation function*) that maps scene $\mathbf{S}$ to an image plus the noise $\mathbf{N}$ (model of error):

$$\mathbf{I} = \Pi(\mathbf{S}) + \mathbf{N} \tag{3.1}$$

The image formation function is often irreversible. For example, a 2-D image of a 3-D scene does not contain the depth information of the scene. Thus we may not be able to uniquely interpret the scene given the image data. In addition, the noise in the process of image formation makes the information provided by the image data unreliable. The degree of certainty of an interpretation can be characterized as $P(\mathbf{S}|\mathbf{I})$, the probability of scene $\mathbf{S}$ conditioned on image observation $\mathbf{I}$, which is called *posterior probability distribution.*

Bayes' rule gives a way to compute the posterior probability, that is

$$P(\mathbf{S}|\mathbf{I}) = \frac{P(\mathbf{I}|\mathbf{S})P(\mathbf{S})}{P(\mathbf{I})} \tag{3.2}$$

For the given image data, $P(\mathbf{I})$ has a fixed value that does not depend on $\mathbf{S}$ and we treat it as normalization constant. Then

$$P(\mathbf{S}|\mathbf{I}) \propto P(\mathbf{I}|\mathbf{S})P(\mathbf{S}) \tag{3.3}$$

where $P(\mathbf{S})$ is the prior probability we mention above. $P(\mathbf{I}|\mathbf{S})$ is referred to as the *likelihood function* for $\mathbf{S}$. $P(\mathbf{I}|\mathbf{S})$ depends on the image formation function Π and the noise $\mathbf{N}$. If the image data is uncorrupted by noise and errors, the image formation function Π will project $\mathbf{S}$ to a unique $\mathbf{I} = I$. For other image $I' \neq I$, $P(I'|\mathbf{S})$ will be zero. However, due to the noise, the resulting image data $\mathbf{I}$ may have a value other than I and we may not know what it will be. In this case, $P(\mathbf{I}|\mathbf{S})$ will be non-zero when $\mathbf{I} \neq I$. The noise has the effect of broadening the distribution of the likelihood function $P(\mathbf{I}|\mathbf{S})$ and makes the information provided by the image data less reliable. For this reason, we may view the likelihood function as a way to explicitly model the noise and errors.

The computation of $P(\mathbf{S}|\mathbf{I})$ from $P(\mathbf{S})$ and $P(\mathbf{I}|\mathbf{S})$ is called Bayesian inference. It provides a way to combine the prior knowledge and obtained data to make inferences about the world. From the above description, we can find the advantages of the Bayesian approach:

- It allows the use of statistical prior knowledge about the world (the prior probability distribution). This effectively constrains the solution space.
- It uses explicit models of noise and errors (the likelihood function). This reflects our knowledge about the process of image formation and the characteristics of the noise.

The Bayesian approach has been successfully applied in many vision problems, including stereopsis [Bar89, GLY95, Bel95, LB95], which we discuss in the following section.

3.2 Bayesian Approach to Stereopsis

In the context of stereo matching, we wish to infer the disparity map $D(x)$ from the given left image $I_l(x)$ and right image $I_r(x)$[1]. Within the Bayesian paradigm, one infers D by considering the posterior probability $P(D|I_l, I_r)$. One approach is to find the most probable D:

$$\hat{D} = arg\max_D P(D|I_l, I_r) \tag{3.4}$$

This is the so called *maximum a posteriori* (MAP) estimate. If we know the prior model $P(D)$ and likelihood function $P(I_l, I_r|D)$, we can apply Bayes' rule to obtain the posterior probability

$$P(D|I_l, I_r) = \frac{P(I_l, I_r|D)P(D)}{P(I_l, I_r)} \tag{3.5}$$

In some approaches, instead of using the probability directly, the following energy functions are used

$$E_S = -log(P(I_l, I_r|D)P(D)) = E_D + E_P \tag{3.6}$$

where $E_D = -\log P(I_l, I_r|D)$, $E_P = -\log P(D)$. The problem now is to choose D such that E_S has minimum energy.

In order to use Bayesian approach in stereo matching, we need to define

- The prior model $P(D)$, where D is the disparity map. The prior model contains assumptions about the scene geometry.
- The likelihood function $P(I_l, I_r|D)$, which reflects the noisiness and ambiguity in the image formation.

[1] For notation convenience, we use I_l and I_r to refer to the left and right images. This does not imply we only use the image intensity as image feature. In fact, we can use other features (such as phase, edges, zero-crossings) extracted from both images

3.2.1 Likelihood Function

The form of the likelihood function depends on the image data. In [GLY95], for example, intensity is used to derive the likelihood function. To simplify the presentation, we assume that there is no occlusion in the scene and that the surfaces in the scene are Lambertian. By assuming additive Gaussian white noise in the images, the likelihood function can be written as

$$P(I_l, I_r|D) = \frac{1}{(2\pi v^2)^{N\times M}} e^{-E_D} \tag{3.7}$$

where

$$E_D = \frac{1}{4v^2}\sum_{x,y}(I_l(x,y) - I_r(x + D(x,y), y))^2 \tag{3.8}$$

where I_l and I_r are the image intensities, v is the standard deviation of the Gaussian noise. $D(x,y)$ is the disparity at location (x,y), $N \times M$ is the size of the image. Although this function is developed under the assumption of Lambertian illumination and uses intensity as the image feature, it can be rewritten by replacing I_l and I_r with general feature function F_l and F_r. F_l and F_r can be any image features that are viewpoint invariant, such as edges, texture, or filter outputs:

$$E_D = \frac{1}{4v^2}\sum_{x,y}(F_l(x,y) - F_r(x + D(x,y), y))^2 \tag{3.9}$$

As we have discussed in previous chapter, phase has many desirable properties and it may be a good candidate for stereo matching purpose. Instead of directly using phase, we use the phase-based binocular measurements described in Section 4.5, as they provide a reliable way to measure the goodness of fit between the phase-shifted left and right signals. The derivation of the likelihood function using the binocular measurements is discussed in Chapter 4.

3.3 MRF as Prior Model

In recent years, Markov random fields (MRF) have been popular prior models of scene structure. The MRF model is an extension of the one-dimensional Markov process. It has attracted much attention in the image processing and computer vision community since the publication of the highly influencial paper by Geman and Geman [GG84]. The main advantage of the MRF model is that it provides a general and natural model for the interaction between spatially related random variables.

We focus our discussion of MRFs in the context of low-level image processing and computer vision. Consider digital images defined on a two-dimensional $M \times N$ pixel lattice. Let D denote the image and let D_{ij} be the pixel at position (i, j). The pixel values may be the intensity of a grey level image, or may be a multivariate value containing the intensity at different wavelengths. In stereopsis, it may also be the disparity at the location of the pixel. A Markov Random Field can be used to describe the global properties of an image in terms of local properties. The local property can be expressed by a conditional distribution

$$P(D_{ij} = d_{ij} | d_{st}, (s,t) \neq (i,j)) = P(D_{ij} = d_{ij} | d_{st} \in R_{ij}, (s,t) \neq (i,j)) \tag{3.10}$$

where R_{ij} is the set of pixels in the neighborhood of pixel at (i, j). The meaning of the above equation is that the distribution of pixel values D_{ij} given the whole image only depends on the pixels in its neighborhood R_{ij}. Fig. 3.1 shows the first and second order neighborhoods that are often used in image processing.

The local interaction between pixels is defined through the energy function $U(R_{ij})$. The joint distribution for a whole image $D = d$ can be expressed as the following Gibbs

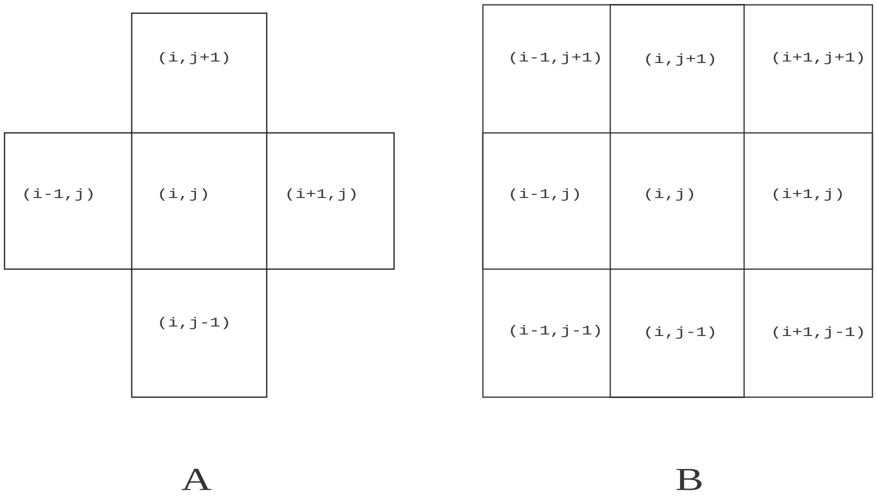

Figure 3.1: A: First order neighborhood. B: Second order neighborhood.

distribution

$$P(D=d) = \frac{1}{Z}\exp(\sum_{i,j} -U(R_{ij})) = \frac{1}{Z}\exp(-V(d)) \tag{3.11}$$

where $V(d)$ is the sum of energies at every pixel in the entire image.

To use the MRF model, the most important task is to define an appropriate energy function $U(x)$. $U(x)$ has a direct impact on the performance of an algorithm in that it contains prior knowledge of the model, which, through Bayes' rule, helps to make decisions from a set of noisy measurements. If the prior knowledge or assumptions of the model do not reflect what the real world is, they will lead to distorted decisions.

One of the commonly used models is the quadratic model, whose energy function is defined as

$$U(R_{ij}) = \lambda \sum_{(k,l)\in R_{ij}} [D(i,j) - D(k,l)]^2 \tag{3.12}$$

The use of the quadratic model in stereopsis can be found in [Bar89, GLY95, Bel95]. The quadratic model imposes a smoothness constraint on the scene geometry. For a smooth surface, the term $[D(x,y)-D(k,l)]^2$ is generally smaller than for a rough surface, which means the smooth surface has a lower energy E_D (defined in Eq. (3.9)) and higher probability.

It is interesting to note that the quadratic model is closely related to the regularization criterion proposed by Poggio, *et al* [PTK85], which is

$$E = \int\int \{[(I_l(x,y) - I_r(x + D(x,y), y))]^2 + \lambda(\nabla D)^2\} dxdy \tag{3.13}$$

where I_l and I_r are intensity functions in the left and right images, D is the disparity map, and ∇D is the gradient of disparity. $(\nabla D)^2$ is interpreted as $\nabla D \cdot \nabla D$, or the square of the magnitude of the disparity gradient. λ is the regularization parameter. The goal of a regularization algorithm for stereopsis is to find a disparity map D such that E has minimum value. The term $\lambda(\nabla D)^2$ serves as the smoothness constraint , where a small λ may result in a noisy solution while large λ may lead to a solution that is over smoothed.

Despite the relation between the Baysian approach and the regularization technique, we argue that the Bayesian approach is more flexible than the standard regularization approach because it can readily be modified to incorporate prior assumptions appropriate for different domains. In the real world, scenes usually have quite different structure. It is hard to capture all the quantities of these structures by only one model. Both the regularization technique and the MRF model using above energy function assume that there is only one smooth and continuous surface in the scene, and the surface changes are small compared to the viewer distance. This is obviously not the case in the real world. Most scenes actually contain several surfaces with the disparity function discontinuous at the object boundary. The surfaces may also have steep slopes and creases. By taking the Bayesian approach, one can easily change the prior model that best fits the structure of the given scene without major modification to the algorithm.

3.4 Piecewise Smooth Functions

So far we have been discussing quadratic energy functions for the prior model. By using this form of prior model, we assume a simple world consisting of only one smooth surface. However most scenes actually contain several surfaces, with disparity discontinuities at the object's boundaries. The quadratic model flattens steeply sloping surfaces, over-smooths surface ceases and over-smooths discontinuities at the object boundaries [Bel95]. In this section, we discuss other possible models that may overcome these problems.

[GLY95] compared several prior models. Although they are 1-D models such that each pixel has only two neighbors, they can be extended to 2-D cases. Besides the quadratic model, there is one that is based on work on visual reconstruction [GG91], whose energy function $U_{\mathrm{e}ff}$ is given by

$$U_{\mathrm{e}ff}(D) = \gamma - \sum_{l} \ln(1 + e^{[\gamma - \mu(D_{l+1} - D_l)^2]}) \tag{3.14}$$

where μ and γ are parameters to be estimated. The third model has an energy function

$$U_{\mathrm{e}ffa}(D) = \mu \sum_{l} \sqrt{|D_{l+1} - D_l|} \tag{3.15}$$

It is argued in [GLY95] that $U_{\mathrm{e}ff}$ encourages a staircase-like disparity function while $U_{\mathrm{e}ffa}$ encourages a single disparity discontinuity. In other words, $U_{\mathrm{e}ff}$ would over-smooth the object boundaries, similar to the quadratic model. Fig. 3.3 shows the

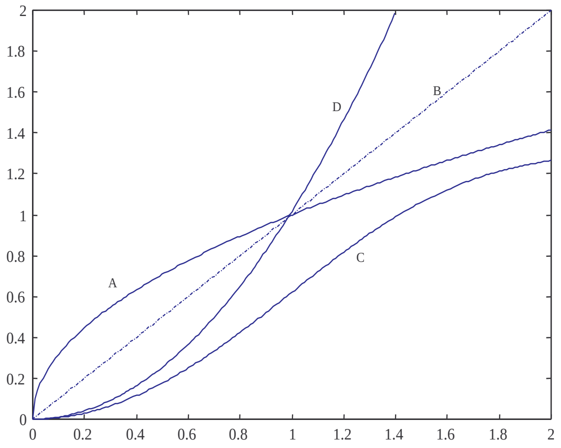

Figure 3.2: Different energy functions. A: $\sqrt{x}$; B: x; C: $\ln(1+e) - \ln(1+e^{1-x^2})$; D: x^2. Here x represents the disparity change between neighboring pixels. One can see that $\sqrt{x}$ has the largest derivative at $x \approx 0$.

effective probability distributions corresponding to the two different models. Note that for the quadratic model, the probability decreases rapidly with the increase of the disparity, which means it assigns very low probability to large disparity difference. This works fine on a smooth surface but fails at the object boundaries, where large disparity differences occur. In order to preserve object boundaries without the effect of over-smoothing, we need a model that tolerates large disparity changes. The distribution of square root model has a longer tail, which means it not only encourages smooth surfaces, but it also allows for large disparity changes. Therefore, the square root model may be a better choice to model piecewise smoothness.

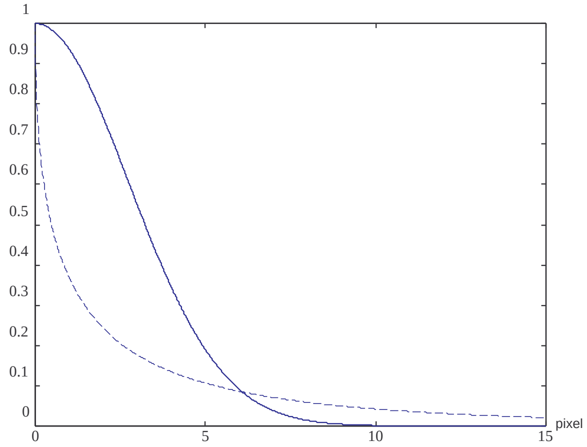

Figure 3.3: The pdf of prior models. Solid line: quadratic model $\exp(-x^2/\mu)$; Dash line: square root model $\exp(-\sqrt{x}/\mu)$. The square root model has a longer tail than the quadratic model, which encourages larger disparity.

3.5 Computational Complexity

Given the prior model and the likelihood function, one can use Eq. (3.5) to obtain the posterior probability for a disparity map D. The Maximum A Posterior Estimate (MAP) can be then used to find the desired disparity map D^*

$$D^* = \arg\max_D P(D|I_l, I_r) \tag{3.16}$$

In practice, this straightforward approach incurs tremendous computational overhead. If the image size is $N \times M$, and the disparity takes the value from $-d_{max}$ to d_{max}, then the Bayesian approach with a MRF prior model requires that we find a disparity map from $(2d_{max})^{N\times M}$ candidate maps that maximizes Eq. (3.16). To obtain the optimal solution by an exhaustive search method, the computational complexity is an exponential $O((2d_{max})^{N\times M})$, which in most cases is computationally prohibitive.

One has turn to sub-optimal solutions, such as *stochastic relaxation* [GG84, Bar89], which is often described as *simulated annealing* due to its conceptual similarity to a physical process called annealing. Appendix B provides details about this algorithm. Simulated annealing is an iterative algorithm, which converges to the desired result after a number of iterations. However, the convergence would be rather slow if the images have large size and the disparity resolution is too fine. In addition, the number of iterations varies with different images. One may be unable to predict the running time of the algorithm. This makes the algorithm impractical for real time application.

In recent years, researchers found that the MRF model can be replaced by a multiscale stochastic model, which can eliminate the iterative procedures while achieving compatible results, with much less computation [BS94, LKWT93]. [LKW94] has developed a theoretical framework to justify the model. More details about the multiscale model are given in Chapter 5.

Chapter 4

Likelihood Function

In this chapter, our task is to find a likelihood function that characterizes the information provided by the image data. In order to derive the likelihood function, we need to explicitly model the sources of variability (uncertainty) for the image data.

The image data we use are the phase-based binocular measurements (Eq. (2.17)) discussed in Section 2.5. Besides the desirable properties of binocular measurements, it is also interesting to note that neurophysiological research has shown that the first stages of disparity processing in the primary visual cortex in cats, primates, and in the visual wulst in owls are thought to use a phase-based measurement. Understanding the statistical properties of binocular measurement not only enables us to use it within the Bayesian framework in computer vision, it also helps us to understand how the measurement might be used in biological vision.

There are several sources of variability of binocular measurements. In this chapter, we identify and model these sources so that we can derive the likelihood function. We investigate how these sources affect the likelihood function in different pre-shifts τ, scales λ, and orientations θ. Finally, we derive a joint likelihood function using the family of binocular measurements at different pre-shifts for a single scale and orientation.

4.1 Variability of Binocular Measurement

Ideally, according to Eq. 2.17, when the left and right signals are perfectly matched locally at a location x_0 with a pre-shift τ, the binocular measurement $C(x_0, \tau)$ should equal one. In practice this is not always the case, owing to:

- *Noise*: The sensor may introduce noise into the images, for example, quantization error, noise in the imaging system, etc.

- *Smooth but non-constant Surface*: The binocular measurements were derived from an assumption of a constant disparity field. However in natural images, this is rarely the case. This results in disparities that are not constant within the Gaussian window used in Eq. 2.17. The left and right images are no longer perfectly shifted version of each other within the window, and it is not possible to bring them into match everywhere with a simple shift. This decreases the cross-correlation of the two images and makes the magnitude of C smaller than one.

- *Variation in the instantaneous frequency*: For measurement of disparities from local phase-differences as explained in Section 2.2, the variation of the instantaneous frequency is a source of measurement error. One can see this from Eq. (2.6), which requires the estimation of the average instantaneous frequency between the left and right signals to compute the shift. The local variation of instantaneous frequency depends on the local image structure.

- *Discontinuities*: Discontinuities not only cause non-constant disparities within the Gaussian window, they also cause occlusion. It is impossible to establish correspondence when occlusion happens. Therefore the information that the binocular measurements provide cannot be used to determine the disparity reliably.
- Deformation/Scale change: The perspective projection may cause significant geometric deformation and contrast variation between left and right views.

Because of these factors, the binocular measurements $C(x, \tau)$ will not always equal one even when the disparity $d(x)$ at location x is equal the pre-shift τ. Given this nature of uncertainty, we may wonder what information we can obtain from the binocular measurements and what behavior we should expect from them. With a Bayesian approach, we may characterize the variability of C with a likelihood function $p(C|D)$. Since C is obtained at different pre-shifts τ, scales λ, and orientations θ, we also want to know how $p(C|D)$ depends on τ, λ and θ, given the disparity. In the remainder of this chapter, we first model several sources of variability and apply these models in real images to derive the likelihood function.

4.2 Models of Variability Sources

Among the sources of variability we mention above, the deformation/scale change and discontinuity are generally difficult to model and we leave it for future research. Here we only model the *noise*, the *non-constant disparities* and the *variation in instantaneous frequency*.

As a first step, we can use real images to model the variation of instantaneous frequency. According to Eq. (2.5), a pure sinusoidal signal has constant instantaneous frequency. The filter outputs of real images are not purely sinusoidal. In this case, the instantaneous frequency depends on the local image structure and is not constant. In the absence of other sources of variability, the two images are simply translated version of each other, and therefore the left image and right image are identical when they are brought to perfect match. When not aligned perfectly, binocular measurements depend on the local behavior of instantaneous frequency. The larger the difference between disparity D and pre-shift τ, the greater the variance of the binocular measurement [FJ93].

The noise is commonly modeled as additive white noise [GLY95]; here, we simply add non-zero, Gaussian white noise to one of the images. In this way, one should expect a higher signal-to-noise ratio (SNR) at coarser scales, because of the difference in the power spectra between natural images and white noise images. Usually, natural images have power that decreases with frequency (e.g., the $1/f^2$-like power spectrum [Cas96]). But white noise does not have a decreasing power spectrum with frequency. Therefore, signal-to-noise ratios typically decrease as frequency increases. This means that the measurements at finer scales are generally expected to be more noisy. Typically, in 8 bit images noise is about $2-4\%$ of the average image intensity. Since most images in our experiment have an average image intensity of about 100, here we assume Gaussian white noise that is mean-zero with a standard deviation of 2 gray levels for 8 bit gray scale images.

To model the non-constant disparity, one needs a model for surface properties. The surfaces of natural scenes can usually be represented by a fractal model [Bel95], that is, a $1/f^2$-like process. Here, to generate a smooth, but non-constant disparity map, we generate a synthetic fractal disparity map, which we refer to as "displacement noise", and then we warp one of the two images accordingly using cubic interpolation. The question is how much displacement noise we should add to properly model this effect. According to Eq. (1.2), the distance from the surface to the image plane can be determined from the disparity d:

$$z = \frac{ft}{d}$$

where f is the camera focal length, t is the distance between two cameras, z is the distance from the surface to the image plane. Assume that the average distance from the surface to the image place is $\bar{z}$, which corresponds to the average disparity $\bar{d}$. A change of distance Δz causes a corresponding change of disparity Δd, where the relation between Δz and Δd can be expressed as

$$\bar{d} \cdot z = (\bar{d} + \Delta d) \cdot (\bar{z} + \Delta z) = ft$$

which can be rewritten as

$$\frac{\bar{d}}{\bar{d} + \Delta d} = \frac{\bar{z} + \Delta z}{\bar{z}} \tag{4.1}$$

The term $\frac{\bar{z}+\Delta z}{\bar{z}}$ represents the "roughness" of the surface, which may be a slope, crease, or stochastic process (Fig. 4.1). It may be reasonable to assume that $\frac{\bar{z}+\Delta z}{\bar{z}}$ is between $90 - 95\%$ if the surface is smooth. Then the choice of Δd depends on the average disparity $\bar{d}$. For example, if the average disparity is 4 pixels, Δd is between $0.2 - 0.4$ pixel. The value of Δd increases with $\bar{d}$. One can have a raw estimation of the average disparity before adding the displacement noise. Normally, the average disparities in our experiments are between $3-6$ pixels and we have used displacement noise with variance of 0.2.

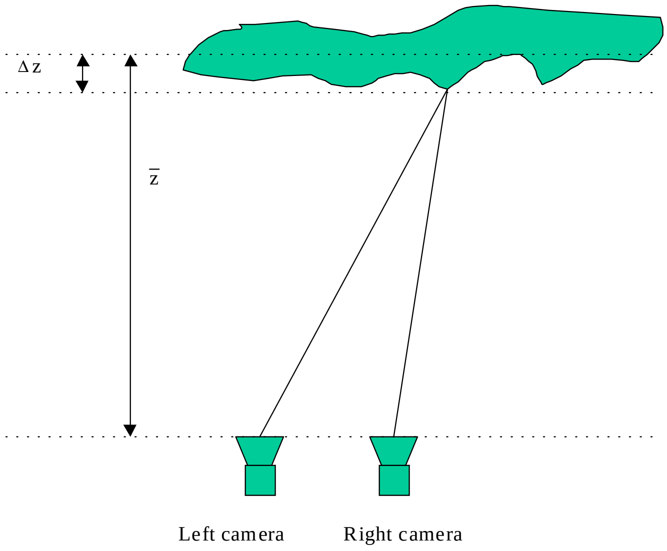

Figure 4.1: The roughness of a surface. $\bar{z}$ is the average distance to the image plane, Δz is the change of the surface. The roughness of the surface can be represented by $\frac{\bar{z}+\Delta z}{\bar{z}}$.

The displacement noise should form a smooth surface without occlusion. Let $D(x)$ be the disparity at location x, according to [GLY95], all surfaces visible to both the left and the right cameras have $|D(x+1) - D(x)| \leq 1$. Therefore, the easiest way to avoid occlusion is to limit the range of disparity between -0.5 and 0.5 pixel.

4.3 Empirical Likelihood Derivation

In this section, we study the effect of these sources of measurement uncertainty on the binocular measurements. Toward this end, we use these models to statistically generate binocular images from which measurement data are collected. From this data, with knowledge of true disparities, we find an empirical likelihood function. Ideally, we want to gather the measurements from many different and independent natural image patches, so that the results fairly reflect the statistical properties of the images in the world. Because the neighborhood pixels in an image are generally highly correlated, it is not feasible to compute the measurements densely over an entire image and use these measurements for statistical analysis. Instead, we should sample many independent patches from the images and obtain one measurement from each patch. To ensure the patches are independent from each other, the size of the patch should be no less than the filter support. For example, if the filter support is 5 pixels, the size of the patch is at least 5×5 pixels.

As the first step of simulation, we use real images as they incorporate natural variations of instantaneous frequency. But for now, we assume no noise and constant disparity $D(x)$, so the left and right images are identical. We can rewrite the pre-shift τ as

$$\tau = D(x) + \Delta\tau \tag{4.2}$$

where $\Delta\tau$ is the distance from the true disparity. In this case, as in all others that follow, we can, without great loss of generality, assume $D(x) = 0$, since the measurements only depend on the difference $\Delta\tau$. Thus the binocular measurement can be obtained using Eq. (2.17) at different values of $\tau = \Delta\tau$. Fig. 4.2 shows the distributions of the real part of the binocular measurements for filters tuned to orientation $\theta = 0^\circ$ and wavelength $\lambda = 9.2$ pixels, at 6 different value of $\Delta\tau$. One may notice that at $\Delta\tau = 0$, the real part of the measurements is equal to 1, a result expected from Eq. (2.17).

The distribution of $\mathrm{R}e[C(x, D(x) + \Delta\tau)]$ begins to spread out when $\Delta\tau$ increases. A wider peak of the curve means the left and right signals are less correlated. When $\Delta\tau$ is large enough, the distribution becomes almost uniform, which means the left

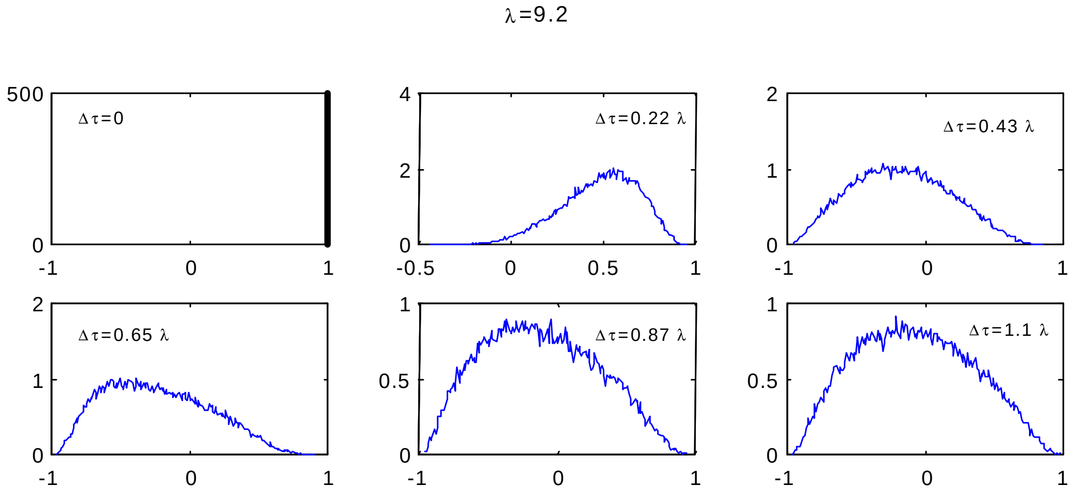

Figure 4.2: The distribution of real part of $C(x, \Delta\tau)$ with different $\Delta\tau$. The filters are tuned to wavelength of $\lambda = 9.2$ pixels with 0° orientation. The only source of variability is the variation of instantaneous frequency.

and right signals are grossly unregistered and the local measurement contains little useful information. The spreading out of the distribution of $\mathrm{Re}[C(x, \Delta\tau)]$ means that its variance generally increases with the increase of $\Delta\tau$ until the distribution becomes almost uniform. Fig. 4.4 shows the curve of the standard deviation of $\mathrm{Re}[C(x, \Delta\tau)]$ as a function of $\Delta\tau$.

From Fig. 4.2, one can also see that, as $\Delta\tau$ increases from 0, the mean value of $\mathrm{Re}[C(x, D(x) + \Delta\tau)]$ gradually decreases, eventually turning negative, then moving back to 0. This phenomena is illustrated in Fig. 4.3, which shows the mean value of the real part of $C(x, \Delta\tau)$ at different $\Delta\tau$. The behavior of the mean value can be explained by Eq. (2.18), which we rewrite here

$$C(x_0, \tau) = C(x_0, D(x) + \Delta\tau) \approx \frac{\sum A_l A_r e^{j\Delta\Phi}}{\sqrt{\sum A_l^2}\sqrt{\sum A_r^2}}$$

The real part of $C(x_0, D(x) + \Delta\tau)$ is

$$\mathrm{R}e[C(x, D(x) + \Delta\tau)] \approx \frac{\sum A_l A_r \cos(\Delta\Phi)}{\sqrt{\sum A_l^2}\sqrt{\sum A_r^2}}$$

When $\Delta\tau = 0$, the left and right signals are in perfect match. Therefore $\Delta\Phi = 0$, which means $\mathrm{R}e[C(x, \Delta\tau)]$ is 1. When $\Delta\tau = 0$ increases, so does the phase difference $\Delta\Phi$. Then the change of $\mathrm{R}e[C(x, \Delta\tau)]$ almost follows the change of the cosine function $\cos(\Delta\Phi)$. However, when $\Delta\tau$ continues to increase, the left and right signals become misaligned to a large degree. The cross-correlation of the signals gradually reduces to zero, which makes the value of $\mathrm{R}e[C(x, \Delta\tau)]$ also close to zero.

From Fig. 4.3, one can find that the wavelength of the mean value curve of $\mathrm{R}e[C(x, \Delta\tau)]$ for the white noise inputs is approximately equal to the filter's tuning wavelength, while the curve for the real image has a longer wavelength. This shows the effect of the instantaneous frequency for the binocular measurement and further supports the use of real images in the simulation. The wavelength of the mean value curve depends on the average instantaneous frequency of the filter output. The instantaneous frequency of the filtered white noise is higher than that of the filtered pink noise. That is because the average instantaneous frequency is equal to the mean Fourier frequency, while the mean Fourier frequency of filtered pink noise (real image) is lower than filtered white noise, since the real image has a $1/f^2$-like power spectrum. Therefore the curve for the real image has a longer wavelength.

For completeness, Fig. 4.6 and Fig. 4.5 show the mean value curve of the imaginary and amplitude component of the binocular measurements, respectively. The imaginery part of $C(x_0, \Delta\tau)$ is

$$\mathrm{I}m[C(x, \Delta\tau)] \approx \frac{\sum A_l A_r \sin(\Delta\Phi)}{\sqrt{\sum A_l^2}\sqrt{\sum A_r^2}}$$

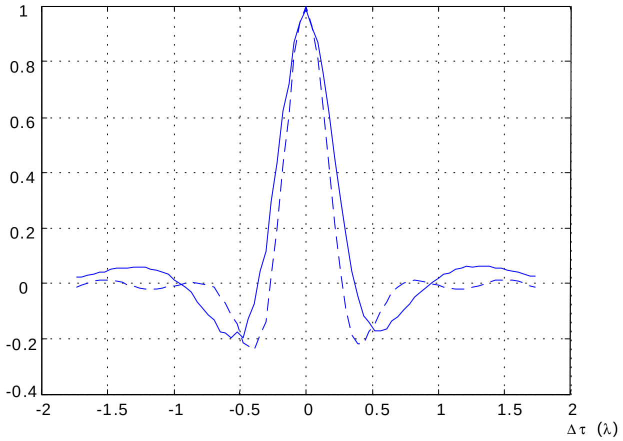

Figure 4.3: An empirical measurement of mean value of real part of $C(x, \Delta\tau)$ vs. $\Delta\tau$. The filter output has a wavelength of about 4.6 pixels. The solid curve is for a typical pair of real images; the dash curve is for synthetic white noise images. The curve for the synthetic image pair has a broader peak than that of real image.

The imaginary part of the binocular measurement is zero when the left and right signals are perfectly matched, because the phase difference $\Delta\Phi$ is zero. When the phase difference $\Delta\Phi$ increases with an increase of $\Delta\tau$, the change of $\mathrm{I}m[C(x, \Delta\tau)]$ should approximate $\sin(\Delta\Phi)$. However, when $\Delta\tau$ continues to increase, the cross-correlation $\sum A_l A_r$ begins to decrease because the two signals become unregistered to a large degree. Therefore, $\mathrm{I}m[C(x, \Delta\tau)]$ is a sinusoid-like curve with a decreasing amplitude as $\Delta\tau$ increases.

The expected value of the amplitude component is the combined result of the real and imaginary part. Fig. 4.7 shows the mean value curve of the phase component of the binocular measurement. One can find that it has a peak value π at about 1/3 of the wavelength λ the filters are tuned to. If the input signals are white noise, we can expect a peak value near $1/2\lambda$, since the average instantaneous frequency of the

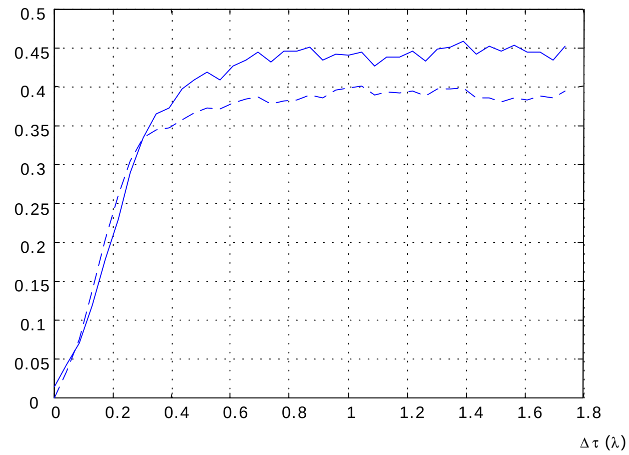

Figure 4.4: The standard deviation σ of the real part $C(x_0, \Delta\tau)$ vs. $\Delta\tau$ for the Gaussian white noise inputs and real image. The solid curve is for a typical pair of real images; the dash curve is for synthetic white noise images. Note that the curve of the white noise image has a steeper slope than that of real image, it is "saturated" approximately 1/4 of the wavelength of the filter output, while the curve of real image is "saturated" at a larger wavelength.

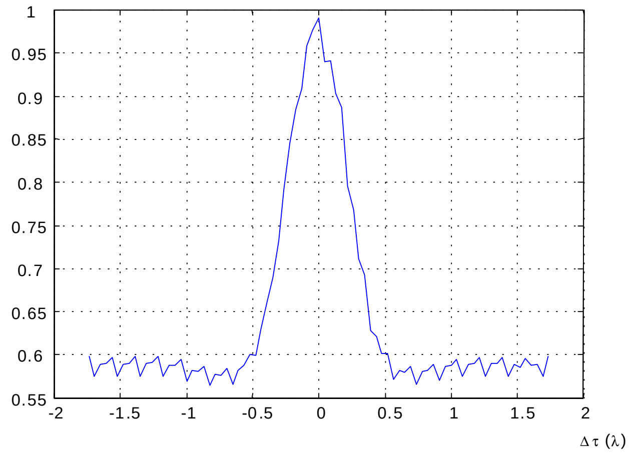

Figure 4.5: An empirical measurement of the expected value of the amplitude component of $C(x, \Delta\tau)$ vs. $\Delta\tau$ for a typical real image pair. The filter output has a wavelength of about 4.6 pixels.

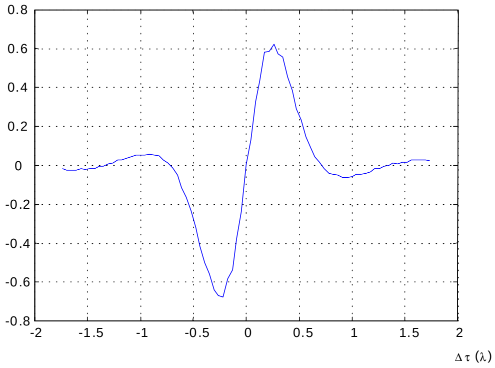

Figure 4.6: An empirical measurement of mean value of imaginary part of $C(x, \Delta\tau)$ vs. $\Delta\tau$ for a typical real image pair. The filter output has a wavelength of about 4.6 pixels.

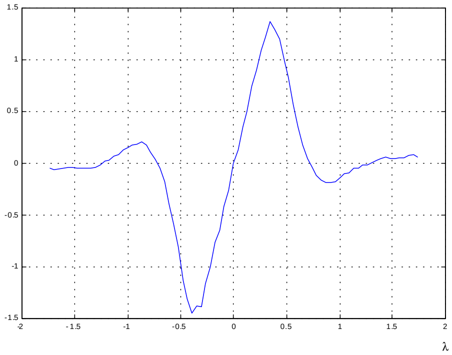

Figure 4.7: An empirical measurement of mean value of phase component of $C(x, \Delta\tau)$ vs. $\Delta\tau$ for a typical real image pair. The filter output has a wavelength of about 4.6 pixels. The curve has peak value π at about 1/3 of the wavelength.

filter output is close to the tuning frequency of the filter. However, in the case of real images, the average instantaneous frequency of the filter output is smaller than the tuning frequency of the filters, due to the $1/f^2$-like power spectrum of real images. Therefore, we should expect a peak between $1/3\lambda$ and $1/4\lambda$ [FJ93].

Until now we have only considered the noiseless case where all variability of C is due to instantaneous frequency. Now, if we take the variation of disparity into account, the left and right images are no longer a translated version of each other. We cannot bring the two images into perfect match everywhere with a single pre-shift. This phenomenon can be simulated by warping one of the images with the synthetic fractal disparity map D_f described in the last section. For a given location x, without pre-shifting the two original images, the true disparity is now $D_f(x)$. To obtain the value of $C(x, \Delta\tau)$, we must use a pre-shift given by $\tau = D_f(x) + \Delta\tau$. Since now the pre-shifts are different at each pixel location, we can no longer bring two images into perfect match with a single constant shift. Therefore, we have to perform the pre-shift individually for each pixel to compute its $C(x, \Delta\tau)$. This is a slow process if the image is large. Fortunately, for the reason we discuss early in this chapter, the computation of the binocular measurements is performed in small patches of the image. It is unnecessary to compute the measurement at one location x using the whole image. We can shift the small patch centered at location x by $D_f(x) + \Delta\tau$ and apply Eq. (2.17) on it. The size of the patch is larger than the Gaussian window in Eq. (2.17) to ensure the independence of the measurements.

Fig. 4.8 shows the distribution of $\mathrm{R}e[C(x, \Delta\tau)]$ for filters tuned to orientation $\theta = 0^\circ$ and wavelength $\lambda = 9.2$ pixels. Note that at $\Delta\tau = 0$, the real part of the measurements are no longer always 1. Instead, it forms a peak near 1 and most the

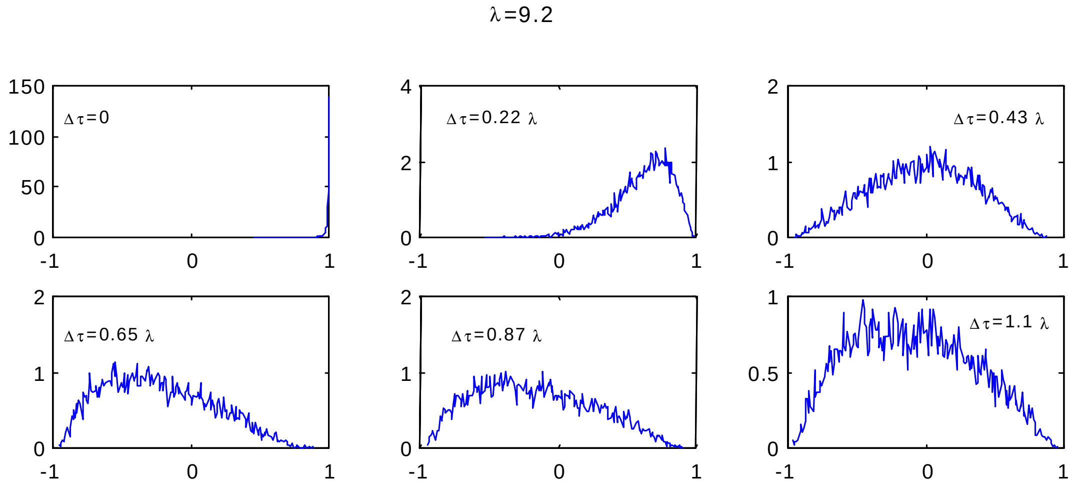

Figure 4.8: The probabilistic distribution of real part of $C(x, \Delta\tau)$ with different $\Delta\tau$. The filters tune to wavelength of $\lambda = 9.2$ pixels with $0°$ orientation. The sources of variability include the variation of instantaneous frequency and the non-constant disparities.

values are between 0.9 and 1. This demonstrates the effect of non-constant disparities on the measurements. Non-constant disparities within the Gaussian window in Eq. (2.17) reduce the cross-correlation between the two local amplitude and phase components of the filter output. Thus, the magnitudes of the measurements will not all have unit magnitude.

Finally, in addition to non-constant disparities and local variations in instantaneous frequency, we now incorporate image noise into the model. As discussed in Section 4.2, one should expect a higher signal-to-noise ratio (SNR) at coarser scales than at finer scales. That means for different scales, the noise has different effects. However, here the measurement variability caused by the image noise is small compared to the effect of instantaneous frequency and disparity noise. Fig. 4.9, Fig. 4.10 and Fig. 4.11 show the distribution of Re$[C(x, \Delta\tau)]$ with filters tuned to orientation $\theta = 0°$ with wavelength $\lambda = 4.6$, 9.2 and 18.4 pixels. Generally, note that the distributions are similar across scales.

Fig. 4.12 shows the measurements with orientation $\theta = 45°$ and filter wavelength $\lambda = 4.6$ pixels. Note that the distributions look different from those at 0° orientation in Fig. 4.10. The reason is that the effective wavelengths that the filters are tuned to, along the epipolar lines, are different (remember that we compute the measurement along the epipolar lines). Assuming the wavelength that the filter is tuned to at orientation 0° is λ, the effective wavelength along the epipolar line is also λ. However, if we steer the same filter to orientation θ, we get an effective wavelength of $\lambda/\cos(\theta)$ along the epipolar line, which is larger than λ. Fig 4.13 shows the mean value curves with the same filter wavelengths but different orientations at 0° and 45°. The curve with $\theta = 45°$ has a longer wavelength than the one with $\theta = 0°$ orientation, because the former has a longer effective filter wavelength.

4.4 Fitting of the Likelihood Function

We have seen the distributions of the measurements at different scales and orientations, with different pre-shifts $\Delta\tau$. These histograms serve as empirical likelihood functions $p(C(x, D + \Delta\tau)|D)$, where D is the disparity. In the above analysis, without loss of generality, we used $D = 0$ to obtain $p(C(x, \Delta\tau|0)$. This is because the curve of $p(C(x, D + \Delta\tau)|D)$ depends only on $\Delta\tau$, which is the difference between the disparity and the pre-shift τ. Since the shape of the distribution curves vary with the value of $\Delta\tau$, it seems that we need to individually fit the curve for each $\Delta\tau$. However, this results in many curves and becomes impractical. It is also difficult to deal with continuous values of $\Delta\tau$ using this approach. Fortunately, although these

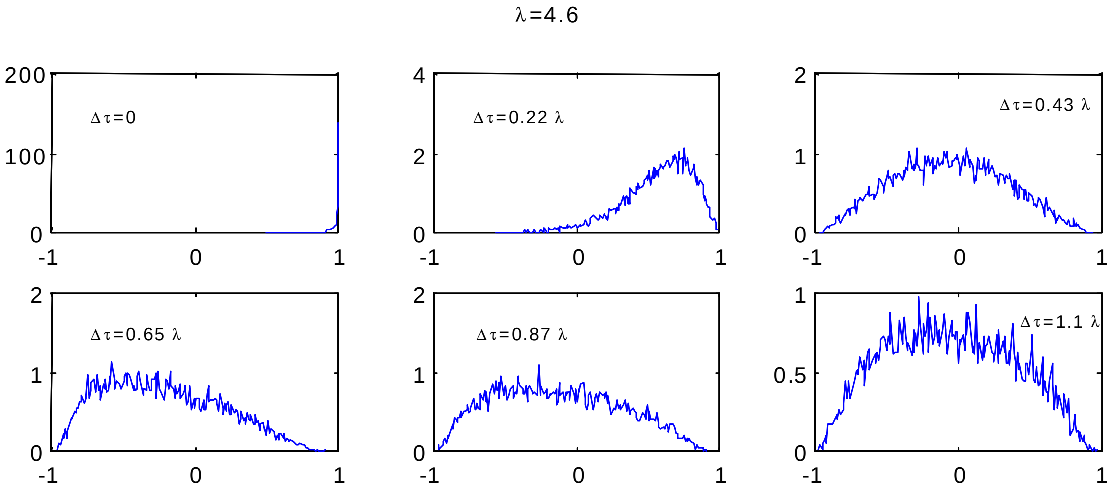

Figure 4.9: The distribution of real part of $C(x, D(x) + \Delta\tau)$ with different $\Delta\tau$ at scale 0. The filters are tuned to a wavelength of $\lambda = 4.6$ pixels with 0° orientation. The sources of variability include the variation of instantaneous frequency, non-constant disparities and noise.

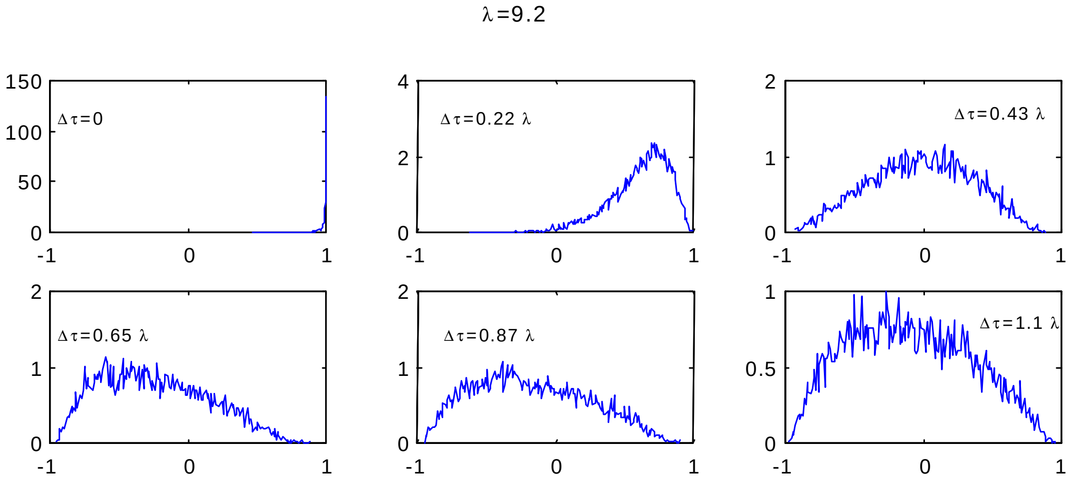

Figure 4.10: The distribution of the real part of $C(x, \Delta\tau)$ with different $\Delta\tau$. The filters are tuned to a wavelength of $\lambda = 9.2$ pixels with 0° orientation. The sources of variability include the variation of instantaneous frequency, non-constant disparities and noise.

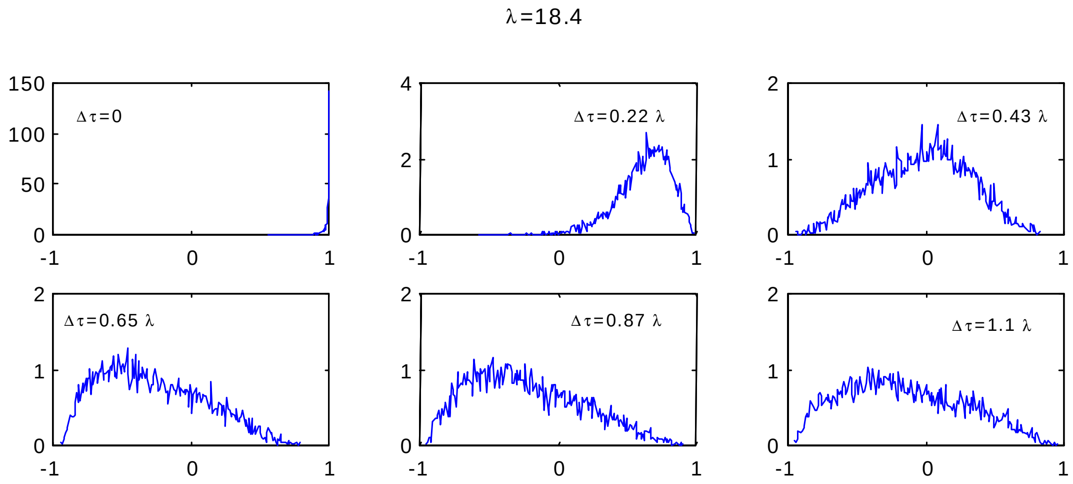

Figure 4.11: The distribution of the real part of $C(x, \Delta\tau)$ with different $\Delta\tau$. The filters are tuned to a wavelength of $\lambda = 18.4$ pixels with 0° orientation. The sources of variability include the variation of instantaneous frequency, non-constant disparities and noise.

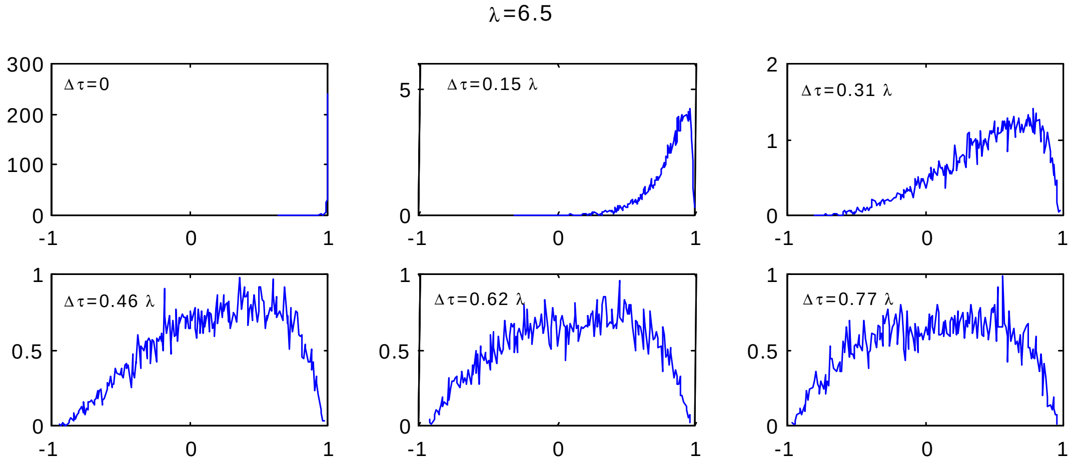

Figure 4.12: The probabilistic distribution of the real part of $C(x, \Delta\tau)$ with different $\Delta\tau$. The filters are tuned to a wavelength of $\lambda = 4.6$ pixels with 45° orientation, with the effective wavelength $\lambda = 6.5$ pixels along the epipolar line. The sources of variability include the variation of instantaneous frequency, non-constant disparities and noise.

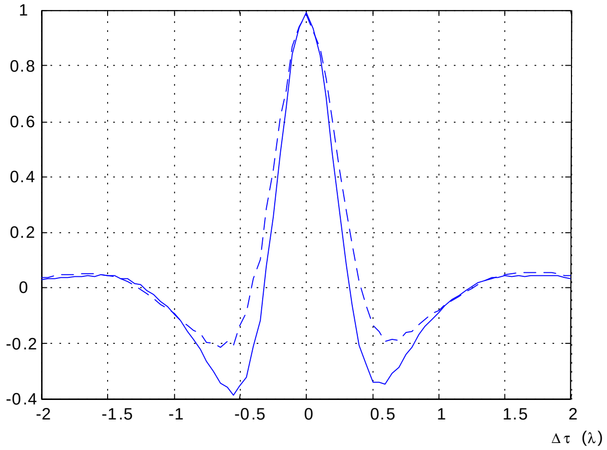

Figure 4.13: An empirical measurement of mean value of the real part of $C(x, \Delta\tau)$ as a function of $\Delta\tau$. The filters are tuned to wavelength of about 4.6 pixels. The solid curve is for a typical pair of real images with filters tuned to 0° orientation; the dash curve is for the same pair of real images with filters tuned to 45° orientation. The curve with 45° orientation has a longer wavelength because the filters have a longer effective wavelength along the epipolar lines.

curves may seem quite different, they can be fitted in the same parameterized forms. More over, the parameters will be functions of $\Delta\tau$, which means that we can describe the behavior of all distributions with only the parameters of the curves and $\Delta\tau$.

The distributions shown in Fig. 4.9- 4.11 can be characterized by a Beta distribution [Muk96]. A random variable is said to have a Beta distribution with parameter a, b ($a > 0, b > 0$), if its probability distribution function is given by

$$f(x;a,b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1, a > 0, b > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

where $B(a,b)$ is the Beta function

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt, \ (a, b > 0) \tag{4.4}$$

Because $C(x_0, \tau_0 \pm \Delta\tau)$ ranges from -1 to 1, Eq. (4.3) can be rewritten as

$$f(x;a,b) = \begin{cases} \frac{1}{2B(a,b)} (\frac{x+1}{2})^{a-1}(1-\frac{x+1}{2})^{b-1}, & -1 < x < 1, a > 0, b > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

To estimate the parameters of the Beta distribution, one can use the method of moments (MM) [Muk96] to obtain a raw estimate, and then use the raw estimate as an initial value to iteratively solve a set of equations derived by the Maximum Likelihood Estimate (MLE) method to get a more accurate estimate. In our experiment, we found that the estimates by the method of moment along are almost as good as the optimal estimates found by the iterative ML method. Therefore, we use method of moment to estimate the parameters. The details are given in Appendix C. Fig. 4.14 shows the Beta distributions fitted to the empirical distribution curves.

In order to see how the distribution changes with the pre-shift, we can examine the curves formed by a and b values at different pre-shifts. Fig. 4.15 to Fig. 4.20 show the a, b values of the Beta distribution as functions of $\Delta\tau$ at different scales

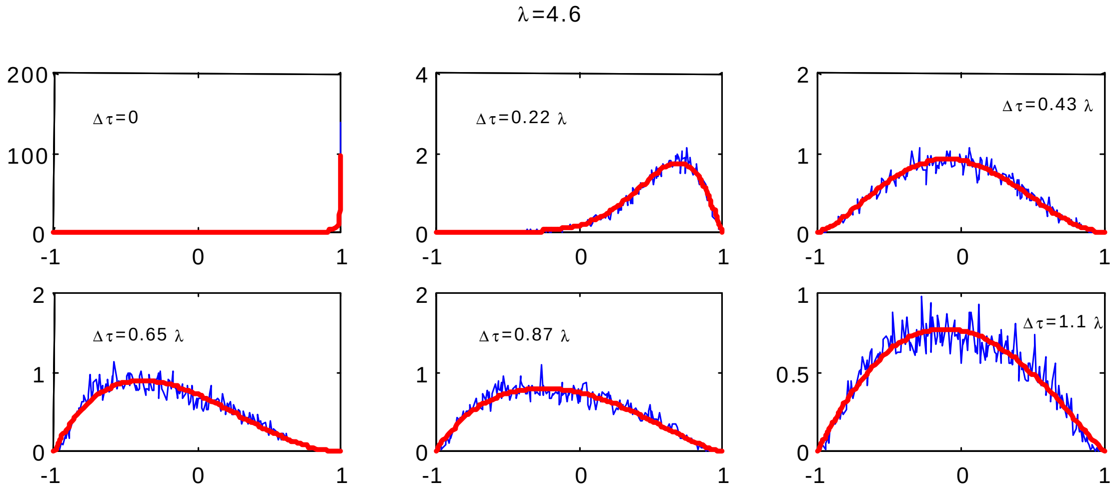

Figure 4.14: Fitting of the Beta distributions to the empirical distribution curves of the six cases shown in Fig. 4.9.

with orientation 0° for three natural images. Only the curves for the positive $\Delta\tau$ are shown. The curves of the negative $\Delta\tau$ are symmetrically related to them. Fig. 4.21 and Fig. 4.22 show the a, b curves at different scales with orientation 45°. One may notice that these curves show similar characteristics for different images and different scales, although in our experiment we use different curves at different scales for higher accuracy.

These curves are useful because with them we can obtain the parameters of Beta distribution at any $\Delta\tau$ through the interpolation of the curves. Let $p(C(x,\tau)|D)$ be the likelihood function for measurement $C(x,\tau)$ at pre-shift τ, then $p(C(x,\tau)|D)$ can be modeled as

$$p(C(x,\tau)|D) = f(C(x,\tau); a(\tau - D), b(\tau - D)) = f(C(x,\tau); a(\Delta\tau), b(\Delta\tau)) \quad (4.6)$$

where $f(C(x,\tau); a(\Delta\tau), b(\Delta\tau - D))$ is the Beta distribution parameterized by a and

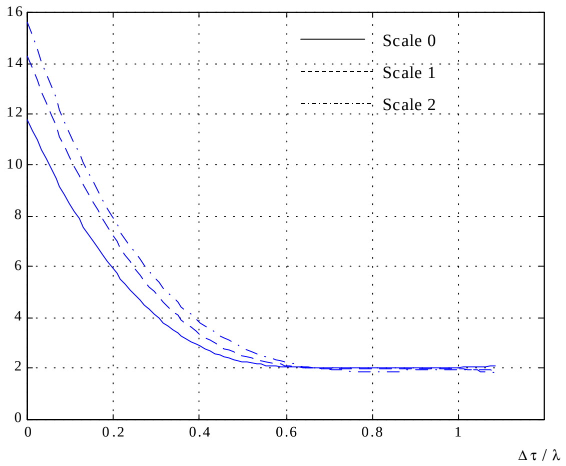

Figure 4.15: The a curve for the "road" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

---

b. $a(\Delta\tau)$ and $b(\Delta\tau)$ are the a and b values at $\Delta\tau$ through the cubic interpolation of the a, b curves.

4.5 Joint Likelihood Function

In the previous section, we derived the likelihood function for the measurement at each individual $\Delta\tau$. In practice, we obtain a sequence of binocular measurements, with different pre-shifts, at each image location. To make better use of the measurements, instead of using a single measurement, we can combine the measurements to estimate the likelihood function. Let $p(C(x)|D)$ denote the joint likelihood function. Unlike the likelihood function $p(C(x,\tau)|D)$ for an individual measurement, there is no τ term in the joint likelihood function because it denotes a family of measurements with different value of τ.

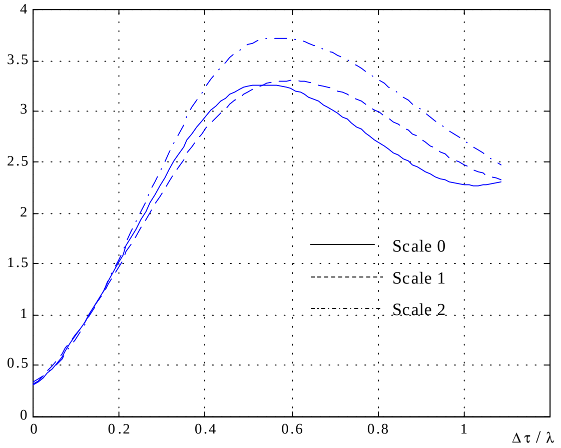

Figure 4.16: The *b* curve for the "road" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

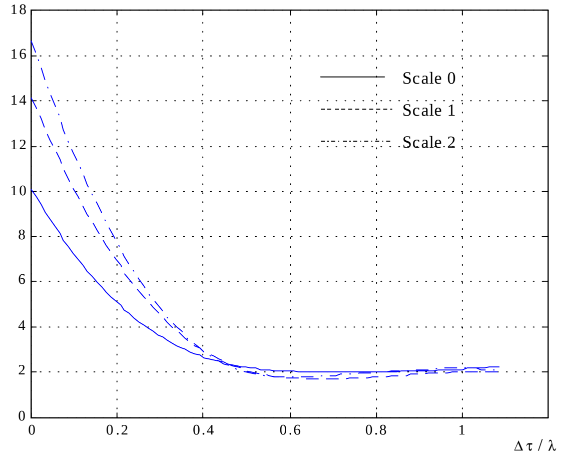

Figure 4.17: The *a* curve for the "wall" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

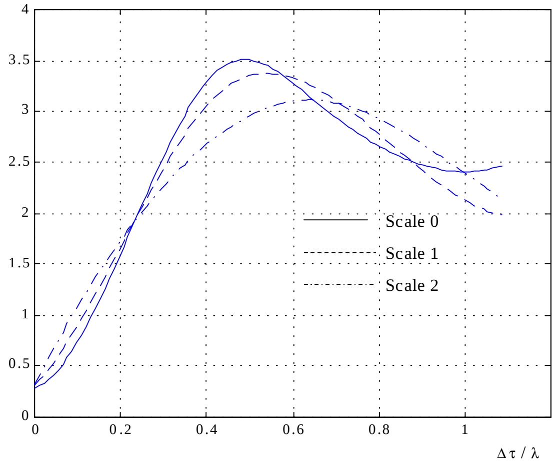

Figure 4.18: The b curve for the "wall" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

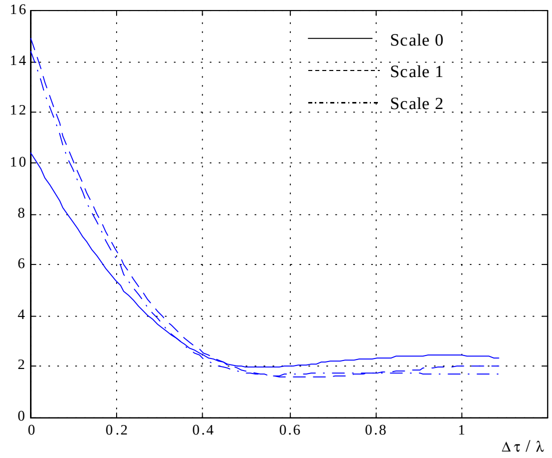

Figure 4.19: The a curve for the "room" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

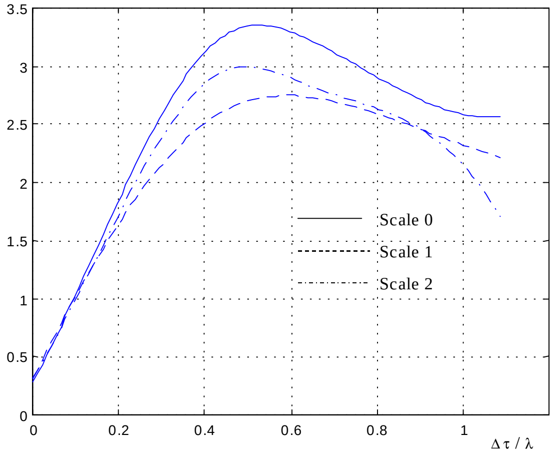

Figure 4.20: The b curve for the "room" image pair. The filters have orientation of 0°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively.

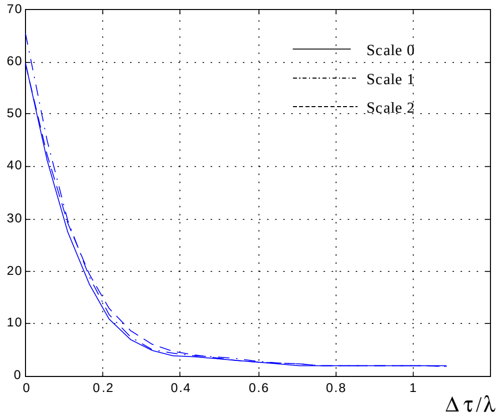

Figure 4.21: The a curve for the "road" image pair. The filters have orientation of 45°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively. The effective wavelength is $\lambda/\cos(45°)$ along the epipolar line, where λ is the filter wavelength.

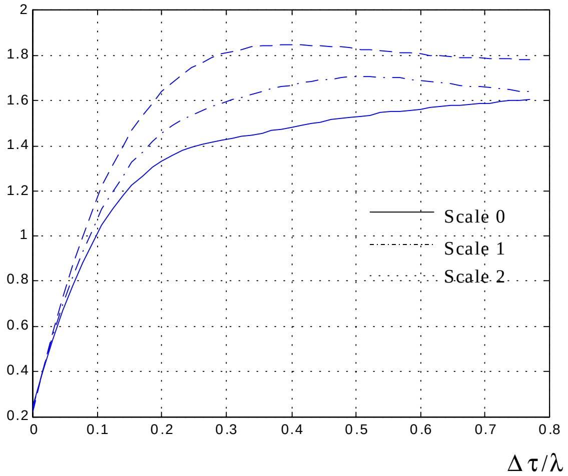

Figure 4.22: The b curve for the "road" image pair. The filters have orientation of 45°. Scale 0, 1 and 2 correspond to filter wavelength of 4.6, 9.2 and 18.4 pixels, respectively. The effective wavelength is $\lambda/\cos(45°)$ along the epipolar line.

The simple way to combine the individual likelihood functions is to take the product of them by assuming independence:

$$P(C(x)|D) = \prod_{\Delta\tau} p(C(x, D + \Delta\tau)|D) \tag{4.7}$$

From Section 4.4, we know that $p(C(x, D+\Delta\tau)|D)$ is Beta distribution parameterized by $\Delta\tau$, a and b, which we rewrite here as

$$p(C(x,\tau)|D) = f(C(x,\tau); a(\tau - D), b(\tau - D)) = f(C(x,\tau); a(\Delta\tau), b(\Delta\tau))$$

Therefore Eq. (4.7) becomes

$$p(C(x)|D) = \prod_{-D_{win} \leq \Delta\tau \leq D_{win}} f(C(x, d + \Delta\tau); a(\Delta\tau), b(\Delta\tau)) \tag{4.8}$$

where $\pm D_{win}$ defines the range of $\Delta\tau$. It is unnecessary to use all the possible values of $\Delta\tau$ because when $\Delta\tau$ is large enough, the probability distribution of $\mathrm{Re}[C(x, D(x)+$

$\Delta\tau)]$ becomes an almost uniform distribution. In practice, we find that choosing d_{win} as one wavelength of filter output achieves satisfactory results.

Sometimes we need to construct a likelihood function at the pre-shift where we do not have binocular measurement. Note that although we obtain the binocular measurements at pre-shift τ with fixed interval Δd, that is

$$\tau \in \{-d_{\max}, \cdots, -2\Delta d, -\Delta d, 0, \Delta d, 2\Delta d, \cdots, d_{\max}\}$$

we can still estimate the likelihood function at any disaprity that is not equal to one of the above pre-shifts. For a disparity $d \in \tau$, we have the measurement $C(x, \tau - D) = C(x, \Delta\tau)$ and we can use Eq. (4.7) to compute the likelihood value. Now consider a disparity $D \notin \tau$, that is, $D = D_0 + \Delta s$, where $D_0 \in \tau$, while Δs represents the sub-pixel accuracy, $0 < \Delta s < \Delta d$. In this case, we do not have the measurement $C(x, \tau - D) = C(x, \Delta\tau + \Delta s)$, since we obtain the measurement at $\tau = D_0 + \Delta\tau$. To estimate the likelihood function with the measurement at $D_0 + \Delta\tau$, we can rewrite Eq. (4.7) as

$$\begin{aligned} p(C(x)|D) &= \prod_{-d_{win} \leq \Delta\tau \leq d_{win}} p(\mathrm{Re}[C(x, D_0 + \Delta\tau)]|D) \\ &= \prod_{-d_{win} \leq \Delta\tau \leq d_{win}} p(C(x, D + \Delta\tau - \Delta s)|D)) \\ &= \prod_{-d_{win} \leq \Delta\tau \leq d_{win}} f(C(x, D + \Delta\tau); a(\Delta\tau - \Delta s), b(\Delta\tau - \Delta s)) \end{aligned} \tag{4.9}$$

$f(C(x, D + \Delta\tau); a(\Delta\tau - \Delta s), b(\Delta\tau - \Delta s))$ is a Beta distribution with parameters a and b. In Section 4.4, we have the curves of a and b with different $\Delta\tau$ values. Because these curves are smooth, one can interpolate the a and b curves to obtain the $a(\Delta\tau - \Delta s)$ and $b(\Delta\tau - \Delta s)$.

The definition of $p(C(x)|D)$ in Eq. (4.8) assumed that the noise in each measurement $\mathrm{Re}[C(x, D + \Delta\tau)]$ is independent at each $\Delta\tau$. However, in fact, the values of $\mathrm{Re}[C(x, D+\Delta\tau)]$ with different $\Delta\tau$ are correlated. The reason for the correlation is the initial linear filters. Even for white noise, and uncorrelated inputs, we expected to see a sinusoidal relation among the values of $Re[C(x, D+\Delta\tau)]$ with different $\Delta\tau$. The curve implies that the values of $\mathrm{Re}[C(x, D+\Delta\tau)]$ near disparity D follow a certain pattern even though they are at different $\Delta\tau$, which invalidates the independence assumption. The dependence of these values means that the product of $p(C(x, D+\Delta\tau)|D)$ will have a sharp peak at the true disparity, or at other pre-shifts where false peaks of binocular measurements occurs (that is, where the pre-shifts are mistaken as true disparity due to the periodic nature of phase). One simple way to overcome this problem is to raise $p(C(x)|D)$ to the power of some value σ. This effectively smooths out $p(C(x)|D)$, which otherwise has very sharp peaks. Fig. 4.23 shows the effect of raising $p(C(x)|D)$ to different values of σ. The value of σ is normally chosen between $\frac{1}{8}$ and $\frac{1}{16}$. We use a value of $\frac{1}{12}$ in our experiments.

4.6 Summary of Results

In this chapter, we developed likelihood functions from binocular measurements. We first identified and modeled the sources of variability in the binocular measurements. Then we empirically derived the form of the likelihood function at different scales and orientations. To make it practical to use the likelihood functions, we fit the form of likelihood functions by a Beta distribution, which is invariant at different scales. Finally, a joint likelihood function is formulated from a sequence of likelihood functions at different pre-shifts for a single scale and orientation.

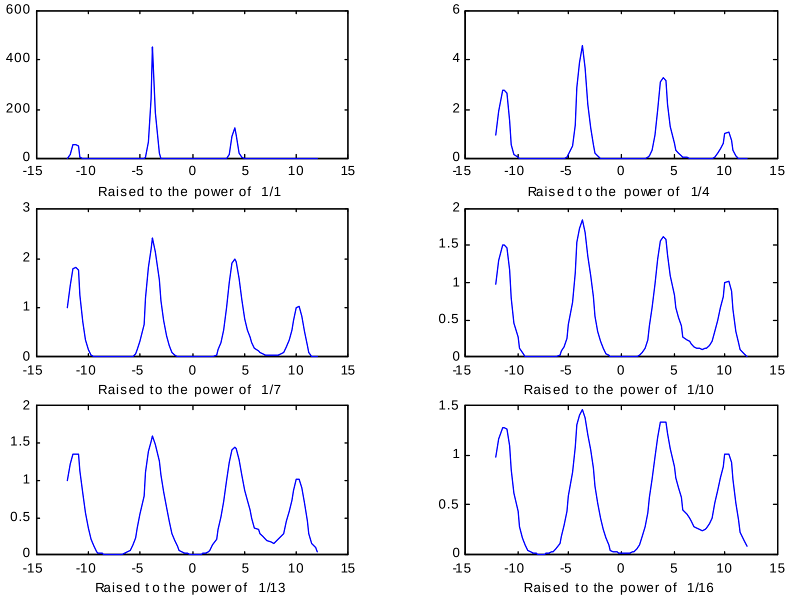

Figure 4.23: Raising $p(C(x)|D)$ to some value σ has the effect of smoothening the curve which otherwise has very sharp peaks. $p(C(x)|D)$ is computed every 0.25 pixel.

Chapter 5

Multi-scale Model for Stereo Matching

In the previous chapter, we discussed how to compute the likelihood function for the observed binocular measurements. Given the likelihood functions at different scales and orientations, we can now investigate how to combine them in an optimal way. This is not only of interest to computer vision, it is also related to biological vision. It has been shown that the first stages of disparity estimate in the primary visual cortex in cats, primates, and in the visual wulst in owls are thought to use the phase-based binocular measurement described in Section 2.4.1. However, the subsequent stages that combine the measurements to find a disparity map are unknown. In [FWH96], Fleet *et al* suggest the linear pooling of the binocular measurements across scales and orientations as the second stage of processing. The primary problem of linear pooling is that it gives every scale and orientation the same weight. It also requires a large number of channels to attenuate false peaks and accentuate true peaks. However, coarse scales have to be interpolated before being added to the fine scales. Since the interpolated results are derived from the known information, the interpolation does not provide any new information. Therefore it may not be appropriate to give the values from the interpolation the same weight as the original values at the fine scales.

The derivation of likelihood functions from the binocular measurements gives us a probabilistic basis for combining the phase-based measurements. It enables us to incorporate the measurements into a Bayesian framework so that we may find an optimal way to combine information across scales and orientations.

In this chapter, we consider two ways of combining measurement information across scales and orientations. First we first try the simplest way, that is, we assume independence of measurements of different scales and orientations. We can then take their products over scales and orientations. As a second approach, we exploit the spatial coherence of typical disparity fields by adopting a multi-scale Markov prior. We show how to correctly propagate information over scales in this multi-scale Bayesian framework. The experimental results are provided for both methods. Our purpose is to show the feasibility of using phase-based measurements in a Bayesian approach. Therefore, we do not expect a significant improvement over other stereo matching methods at this early stage. Although our algorithms seem to work well, there is no in-depth comparison of our results with those from other existing methods. However, the methods we discuss here show potential for further improvement.

5.1 Direct Pooling

If we assume that the measurements are independent at different scales and orientations, and we adopt an uninformative prior (e.g. ignoring smoothness assumption), we come up with a simplest way to combine the information. This involves the multiplication of the likelihood functions across scales and orientations. This approach is somewhat similar to the linear pooling of measurements in [FWH96] and may seem to be nothing significant. However, we claim that the new approach has some advantage over the old one, because it is constructed step by step, with the underlying assumptions being made explicit. Recall that we first identify the sources of variability of binocular measurements and have explicit models for the sources. We then derive the likelihood functions using the models of measurement variability. Finally, with the independence assumption, we multiply the likelihood functions as a way to combine the information. By building the model in this way, we are able to isolate the effects of individual assumptions and test them through experiments. This is particularly important for a biological vision model because we may be able to verify the correctness of the assumptions with neurophysiological data. Therefore, the multiplication model may be a better alternative to the summation model.

The multiplication method can be regarded as a Bayesian approach with uninformative prior and the independence assumption of noise at different scales and orientations. We already know the joint likelihood function $p(C(\tau, x, \theta, \lambda)|D)$ at some pre-shift τ, location x, scale λ and orientation θ. The Bayesian approach requires us to obtain the posterior probability in the form of $p(D|\bar{C})$, where $\bar{C}$ is the set of all measurements. By using the Bayes' rule introduced in Chapter 3, we have

$$p(D|\bar{C}) \propto p(C(\tau, x, \theta, \lambda)|D)p(D)$$

By the assumption of uninformative prior, $p(D)$ is constant. And by the independence assumption of noise at different scale and orientations, we can multiply the likelihood functions across scales and orientations. This results in

$$p(D|\bar{C}) \propto \prod_{\theta_i} \prod_{\lambda_j} p(C(\tau, x, \theta, \lambda)|D) \tag{5.1}$$

In order to multiply the joint likelihood functions, we need to construct the joint likelihood functions so that they have same number of disparities across scale and orientations. The range of disparity is smaller at coarser scale. Therefore fewer measurements are obtained at an image location at coarse scale. We can use the technique in Section 4.5 to construct the likelihood function where no measurement is obtained, so that the likelihood functions at all scales have same number of disparities. The size of the image is also smaller at coarser scale, which results in fewer likelihood functions at coarser scale. This can be overcome by the linear interpolation of the likelihood functions, so that all scales have the same number of likelihood functions.

Once we get the posterior distributions, we have several ways to find the disparity map, such as MAP (Maximum A Posterior estimate) introduced in Section 3.2, and MPM (Maximum Posterior Marginal) [MMP87]. Here we use MAP, which is

$$\hat{D} = arg \max_{D} P(D|\bar{C})$$

However, since the posterior distribution is multi-modal, that is, there may be more than one peak in the distribution, simply choosing the disparity with the largest peak may lose some useful information. One way to use make better use of the posterior probability is to produce a confidence map, which is a histogram of the values of the posterior. Given a disparity with its corresponding posterior value, a broad peak in the histogram near the value means a low confidence for that disparity, while a sharp peak means a high confidence for that disparity.

5.2 Experimental Results

We have implemented the direct pooling method as follows. Similar to the local weighted phase-correlated method, we construct a three-scale Gaussian pyramid from the original images, sub-sampled at each level by a factor of 2 horizontally and vertically. Three quadrature-pair filters are applied at each scale, tuned to orientations 0^o, $+45^o$, and -45^o. The binocular measurements are obtained at every scale and orientation. Then we compute the likelihood function and joint likelihood functions using Eq. (4.6) and Eq. (4.9). Finally, the joint likelihood functions are combined using Eq. (5.1) to get the posterior probabilities. We find the disparity map using MAP.

Some disparity estimates are shown in Fig. 5.1, Fig. 5.3 and Fig. 5.4. Fig. 5.1 is the disparity map of a standard image pair of the Pentagon building as seen from the air (see Fig. 2.5). Fig. 5.3 is the disparity map of the lamp/head sequence (Fig. 5.2), which has a disparity range up to 12 pixels. Fig. 5.4 shows the disparity estimates using frame 2 and 4 from the SRI tree sequence, which is also used to test the local weighted phase-correlation method in Section 2.5. For the SRI tree sequence, compared to the local weight phase-correlation method, the algorithm described here reveals more details of the depth information without making the estimation noisier. The results indicate that the multiplication of the likelihood functions across scales and orientations can server as a better alternative to the direct summation of measurements.

The disparity maps are computed with half pixel resolution. To achieve higher sub-pixel resolution, one can use these disparity maps as raw estimations, and then construct likelihood functions at a smaller sub-pixel interval near the raw disparity

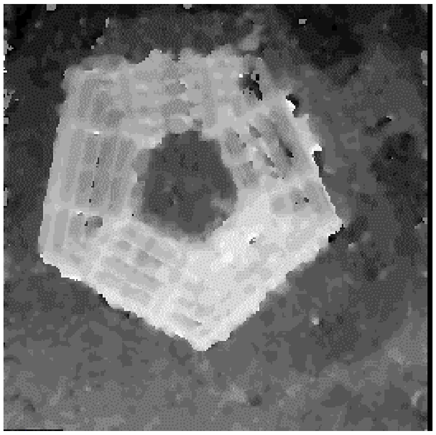

Figure 5.1: Disparity estimate of Pentagon image pair using the product of likelihood functions across scales/orientations. The disparity map has half pixel resolution.

to find a peak, which has higher resolution. This process can be repeated until the desired resolution is achieved.

5.3 Multi-scale Stochastic Model

The direct pooling of measurements in the previous section assumes uninformative prior of the disparity fields. However, disparity fields of natural scenes often have some kind of the spatial coherence, which can be used to obtain a more reliable estimation method. As introduced in Chapter 3, the MRF prior is commonly used for this purpose. In the multi-scale approach, we can combine information across scales using a multi-scale MRF prior within a Bayesian framework. The multi-scale models have been used in a broad range of problems such as image segmentation [BS94], optical flow estimation [LKW94] and classification of texture [CC85], etc. In [LKWT93], it has been shown that the multi-scale models can be used to exactly or approximately

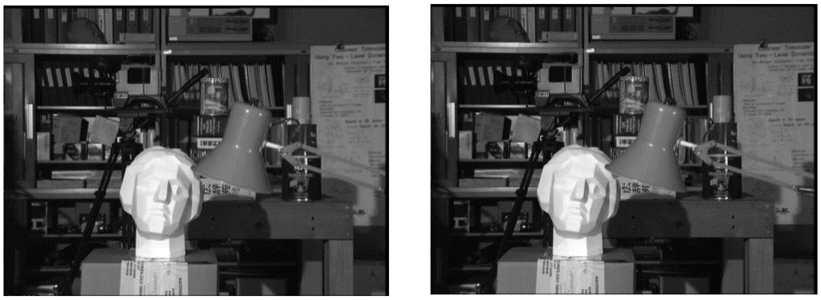

Figure 5.2: The lamp/head sequence.

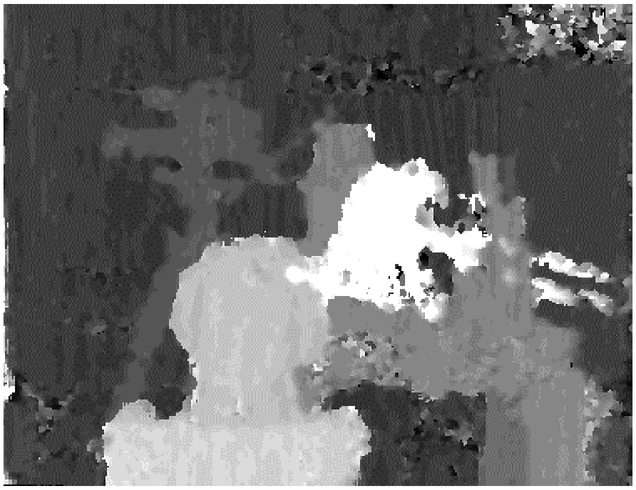

Figure 5.3: Disparity estimate of lamp/head sequence using the product of likelihood functions across scales/orientations. The disparity map has half pixel resolution.

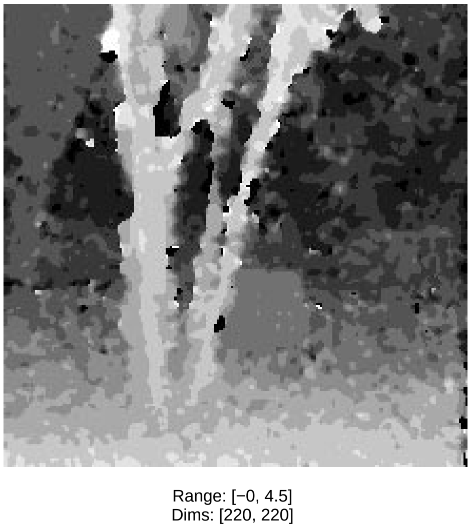

Figure 5.4: Disparity estimate of the SRI tree sequence (frame 2 and frame 4) using the product of likelihood functions across scales/orientations. The disparity map has half pixel resolution.

represent MRFs. Generally, the multi-scale model is composed of a series of random fields progressing from coarse to fine scale. Each field is assumed to only depend on the previous coarser scale, and points in each field are conditionally independent given their coarser scale neighbors. Therefore, the series of fields form a Markov tree in scale. This leads to a scale-recursive model [LKWT93] with computationally tractable properties.

In the case of 2-D signals, the multi-scale model is often described by the *qudatree* shown in Fig. 5.5 in which each node has four descendants. Different levels of the tree correspond to different scale of the fields. For notation convenience, given a node in the tree at a location s, let $s\alpha$ be the location of its parent node, $s\beta$ the location of its child node. For a disparity D_s at node s, it can be shown [LKWT93, LD93] that

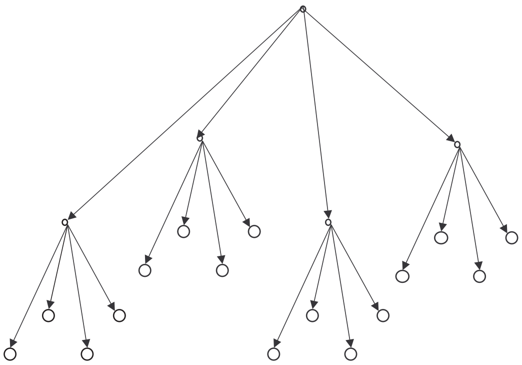

Figure 5.5: The multi-scale model is often described by the *qudatree* in which each node has four descendants.

its relation with its parent nodes, that is, the evolution of the process from coarse scale to fine scales, can be expressed by the scale-recursive multi-scale model:

$$D_s = AD_{s\alpha} + B\omega \tag{5.2}$$

where ω is an independent, zero-mean white noise process, and B is the magnitude of the white noise process. The term $AD_{s\alpha}$ represents the interpolation of the coarse scale to match the sampling grid of the fine scale, where A depends on the particular application and process being modeled. For example, $A = 2$ if the scene property is disparity because the disparity is linearly scaled by the ratio of down sampling. We always assume $A = 2$ as what follows. $B\omega$ represents new information added as the process evolves from one scale to the next.

With Eq. (5.2), the MRF can be transformed into a Markov chain over scales. The Markov chains are formed by the parent and child nodes in the quadtree. It can be expressed as [IB96]

$$p(D_{s\alpha}|D_s) = Z\exp(-\frac{1}{2}B^{-1}(D_{s\alpha} - 2D_s)^2) \tag{5.3}$$

$$p(D_s|D_{s\alpha}) = Z\exp(-\frac{1}{2}B^{-1}(D_{s\alpha} - 2D_s)^2) \tag{5.4}$$

where Z is the normalization constant. The above equations define the transition probability functions for the Markov chain over scales. This model represents the quadratic model discussed in Section 3.4 in multiple scales. It is not surprising that it suffers the same weakness as the quadratic model at a single scale. From Eq. (5.3), it is easy to see that the transition probability decreases rapidly with the increase of the disparity, which means it assigns very low probability to large disparity difference. This works well on smooth surfaces but fails on the object boundaries, where large disparity differences occur.

In order to preserve the object boundary without the effect of over-smoothness, we need a model that tolerates occasional large disparity changes. The square root model discussed in Section 3.4 may be a better choice because it encourages piecewise smoothness while it allows the existence of large disparity difference. To see this, recall that the curve for this model in Fig. 3.3 has a longer tail than the quadratic model. Therefore, in the multi-scale model, we can replace Eq. (5.3) with a more ”robust” model

$$p(D_{s\alpha}|D_s) = Z\exp(-B^{-1}\sqrt{(D_{s\alpha} - 2D_s)}) \tag{5.5}$$

$$p(D_s|D_{s\alpha}) = Z\exp(-B^{-1}\sqrt{(D_{s\alpha} - 2D_s)}) \tag{5.6}$$

where Z is the normalization constant.

Similar to the MRF model, with the multi-scale model, we need to compute the posterior probabilities combining measurements at all scales and orientations, while taking the prior into account. In the next section, we show how to accomplish this using Bayesian belief propagation over scale, to yield an algorithm that is both simple and efficient.

5.4 Belief Propagation

The graph in Fig. 5.5, which represents the multi-scale model described above, is a *Bayesian network* [Wei99, Wei97, Pea88]. In general, a Bayesian Network represents statistical dependencies of variables by a graph. The representation consists of nodes that correspond to random variables and, roughly speaking, arcs that correspond to probabilistic dependencies between the variables. More precisely, the lack of arcs between two nodes represents conditional independence. Fig. 5.6 shows a simple Bayesian network with six nodes. Three of them have observed variables and the other three have hidden variables.

If a Bayesian network is singly connected (i.e., a network without loops), there exist efficient local information passing schemes to calculate the posterior probability at each node. [Pea88] derived a such scheme for a singly connected Bayesian network and showed that the algorithm is guaranteed to converge to the correct posterior probability. In order to propagate information correctly, it is important to avoid "double counting", a situation in which the same information is passed around the network multiple times and mistaken for new information. However, for networks with loops, the propagation of information is much harder because the existence of loops may cause the information to circulate around the loops. As an example, the conventional MRFs are Bayesian networks that contain loops. This explains why it is hard to use them for inference. The convergence is often very slow and may even converge incorrectly if an improper propagation scheme is used.

The multi-scale Markov prior model used here is a tree model without loops, which

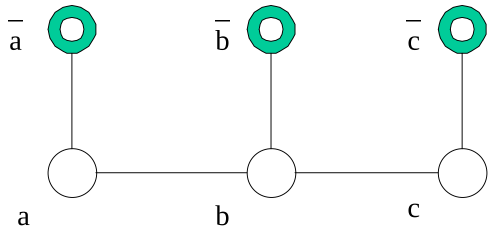

Figure 5.6: A Bayesian network with six nodes, with three observed nodes ($\bar{a}$, $\bar{b}$ and $\bar{c}$) and three hidden nodes (a, b and c).

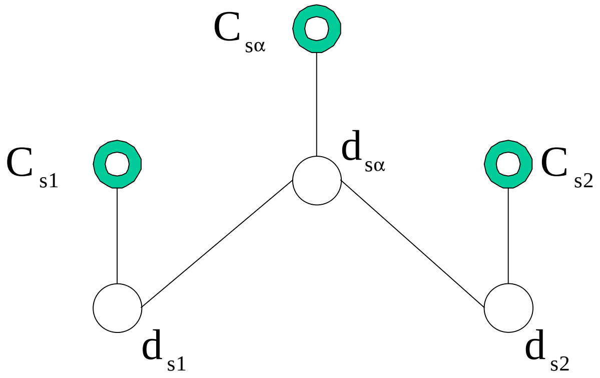

Figure 5.7: A Bayesian network with six nodes. The observed nodes, C_{s1}, C_{s2} and C_{s3}, are the binocular measurements. The hidden nodes, d_{s1}, d_{s2} and d_{s3} are the disparities.

implies the existence of a simple propagation scheme. To derive the correct form of belief propagation, we begin with a simple graph as shown in Fig. 5.6. The variables of the hidden nodes are a, b and c, for which there are corresponding observations $\bar{a}$, $\bar{b}$ and $\bar{c}$ respectively. Let us suppose, for instructional purpose, our goal is to compute the posterior probability $p(b|\bar{a},\bar{b},\bar{c})$. From the graph, because of the conditional independence, the joint distribution $p(a,b,c,\bar{a},\bar{b},\bar{c})$ can be factored into

$$p(a,b,c,\bar{a},\bar{b},\bar{c}) = p(b|a)p(a|\bar{a})p(\bar{a})p(b|c)p(c|\bar{c})p(\bar{c})p(\bar{b}|b)$$

Since $p(a|\bar{a})p(\bar{a}) = p(\bar{a}|a)p(a)$, $p(c|\bar{c})p(\bar{c}) = p(\bar{c}|c)p(c)$, the above equation can also be written as

$$p(a,b,c,\bar{a},\bar{b},\bar{c}) = p(b|a)p(\bar{a}|a)p(a)p(b|c)p(\bar{c}|c)p(c)p(\bar{b}|b)$$

Then

$$\begin{aligned} p(b|\bar{a},\bar{b},\bar{c}) &= \int_c\int_a \frac{p(a,b,c,\bar{a},\bar{b},\bar{c})}{p(\bar{a},\bar{b},\bar{c})}\mathrm{d}a\mathrm{d}c \\ &= kp(b|\bar{b})\int_a p(b|a)p(\bar{a}|a)p(a)\mathrm{d}a\int_c p(b|c)p(\bar{c}|c)p(c)\mathrm{d}c \\ &= kp(b|\bar{a})p(b|\bar{b})p(b|\bar{c}) \end{aligned}$$

where k is a normalization constant that does not depend on a, b and c, and

$$\begin{aligned} p(b|\bar{a}) &= \int_a p(b|a)p(\bar{a}|a)p(a)\mathrm{d}a \\ p(b|\bar{c}) &= \int_c p(b|c)p(\bar{c}|c)p(c)\mathrm{d}c \end{aligned}$$

We can apply a similar technique to the disparity estimation to obtain the propagation rule from the child nodes to parent node (Fig. 5.7)

$$p(D_{s\alpha}|C_{s1},C_{s2},C_{s\alpha}) = kp(D_{s\alpha}|C_{s1})p(C_{s\alpha}|D_{s\alpha})p(D_{s\alpha}|C_{s2}) \tag{5.7}$$

where

$$p(D_{s\alpha}|C_{s1}) = \sum_{D_{s1}\in -D_{max},\ldots,D_{max}} p(D_{s\alpha}|D_{s1})p(C_{s1}|D_{s1})p(D_{s1}) \tag{5.8}$$

$$p(D_{s\alpha}|C_{s2}) = \sum_{D_{s2} \in -D_{max},...,D_{max}} p(D_{s\alpha}|D_{s2})p(C_{s2}|D_{s2})p(D_{s2}) \qquad (5.9)$$

C_{s1}, C_{s2} and $C_{s\alpha}$ are the binocular measurement at node $s1$,$s2$ and $s\alpha$ respectively. $D_{s\alpha}$, D_{s1} and D_{s2} are the disparities. $p(D_{s1})$and $p(D_{s2})$ are the prior probabilities of D_{s1} and D_{s2}. $p(C_{s1}|D_{s1})$, $p(C_{s2}|D_{s2})$ and $p(C_{s\alpha}|D_{s\alpha})$ are the likelihood functions obtained using techniques discussed in Chapter 4. There are several possible choices for $p(D)$. In the case of the human vision system, when people observe a target, normally they will verge the eyes to the target, so that the disparity field will have minimum value, that is, close to zero. Therefore, we can assume that the disparity field is a Gaussian distribution with zero mean. However, for a given pair of images taken from camera, the mean value of the disparity field may not be zero and can be any value in the range of the disparity. In this case, we can assume a uniform distribution of disparity, that is, the disparity at each pixel location is equally likely to be any one of the values in the disparity range $-D_{max}, ..., D_{max}$, Eq. (5.8) and Eq. (5.9) become

$$p(D_{s\alpha}|C_{s1}) = \sum_{D_{s1} \in -D_{max},...,D_{max}} p(D_{s\alpha}|D_{s1})p(C_{s1}|D_{s1}) \qquad (5.10)$$

$$p(D_{s\alpha}|C_{s2}) = \sum_{D_{s2} \in -D_{max},...,D_{max}} p(D_{s\alpha}|D_{s2})p(C_{s2}|D_{s2}) \qquad (5.11)$$

$p(D_{s\alpha}|C_{s1})$ and $p(D_{s\alpha}|C_{s2})$ are called *prediction probabilities* which are, in general, computed from the posterior $p(C_{s1}|D_{s1})p(D_{s1})$ and $p(C_{s2}|D_{s2})p(D_{s1})$ with the transition probabilities $p(D_{s\alpha}|D_{s1})$ and $p(D_{s\alpha}|D_{s2})$. In the case that D_{s1} and D_{s2} are uniform distributions, they can also be computed directly from likelihood functions $p(C_{s1}|D_{s1})$ and $p(C_{s2}|D_{s2})$.

For notational convenience, let E_{st} denote the prediction probability distribution from node s to t, which is computed from the posterior probability at node s. Let

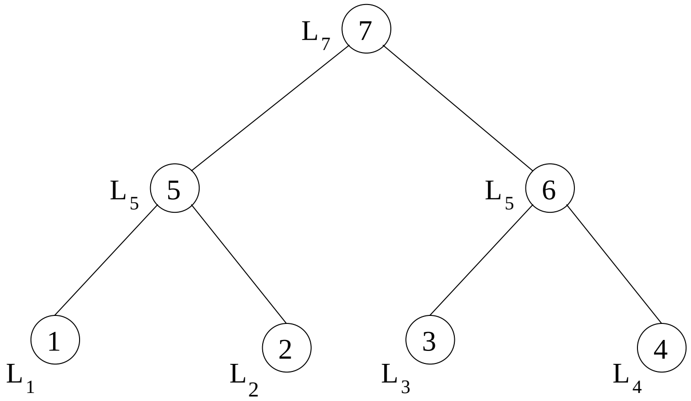

Figure 5.8: A Bayesian network, which is a simple dyadic tree. Each node has the likelihood function pre-computed.

L_s denote the likelihood function at s, and let Q_s denote the posterior probability distribution at s. Eq. (5.7) can then be written as

$$Q_{s\alpha} = E_{s_1 s_\alpha} L_{s\alpha} E_{s_2 s_\alpha} \tag{5.12}$$

Now we can apply the propagation rule to the quadtree model. To simplify the presentation, we base our discussion on the dyadic tree as shown in Fig 5.8. The result can be easily extended to the quadtree model. In Fig 5.8, each node has the likelihood function already computed (i.e. L_1,L_2,...L_6). With the transition probability known, the propagation on the tree can be divided into two phases:

- Upward propagation
- Downward propagation

When designing the information propagation scheme, in order to avoid double counting, it is important to ensure that no nodes receive their own information. That is, their own likelihood functions should not be passed back to them. Therefore, we may need a mechanism that keeps track of the flow of information in the graph. One way to achieve this is to keep a copy of the prediction probability from the child node (which encodes the information from the child node) at each node, instead of direct multiplication with the likelihood function. In this way, we know which part of information comes from which node during the downward propagation.

The details of scheme are described below: We first use the simple dyadic tree shown in Fig. 5.8 as example. The upward propagation is shown in Fig. 5.9. At node 5, we obtain the prediction probability distribution E_{15} and E_{25} from nodes 1 and 2. Instead of multiplying them with L_5, we keep a copy of E_{15} and E_{25} at node 5. At node 7, the prediction probabilities E_{57} and E_{67} are computed from $E_{15}L_5E_{25}$ and $E_{36}E_6E_{46}$ using Eq. (5.10).

The downward propagation is where we should pay attention to double counting. When passing information from node 7 to node 5, the prediction probability E_{75} is computed from $E_{67}L_7$. Note that it should not be computed from $E_{57}L_7E_{67}$, The reason is that E_{57} contains the information from node 5 and we should not send it back to node 5! Similarly, when we pass information from node 5 to node 1, the prediction probability E_{51} should be computed from $E_{25}L_5E_{75}$. It is easy to see that when the posterior probability at node 1 is updated to L_1E_{51}, it contains information from node 2 to node 7. None of them are counted twice.

The complete propagation scheme is summarized in the following:

1. Compute the likelihood functions L at each node of the tree, using techniques discussed in Chapter 4.

2. Upward propagation: Let s be a node in the tree, let R_s be the set of nodes

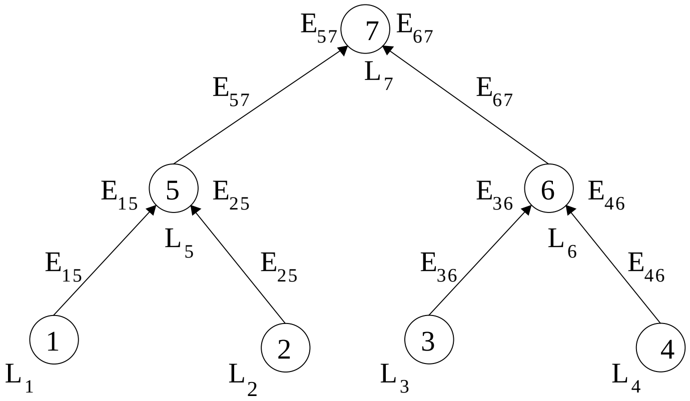

Figure 5.9: A upward propagation in the dyadic tree, with the double counting being taken into account. The prediction probabilities are stored in each node, instead of direct multiplication.

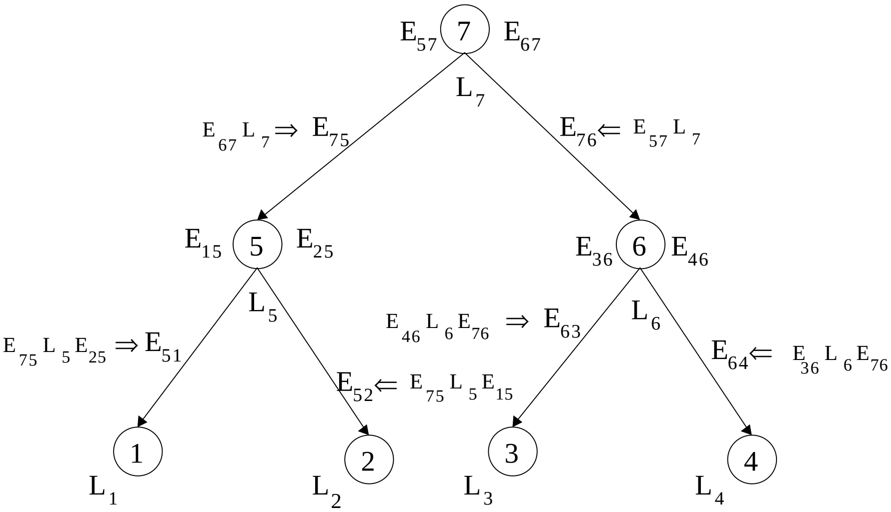

Figure 5.10: A downward propagation in the dyadic tree, with the double counting being taken into account. ⇒ denotes the process of computing the prediction probability from the posterior probability, given the transition probability. The prediction probabilities are stored in each node, instead of direct multiplication. This downward propagation scheme is such that no node receives its own information.

that are children of node s. Compute the prediction probabilities E_{ts} for all nodes $t \in R_s$ using Eq. 5.10, and store the results in node s.

3. Downward propagation: Let s be the parent of node t. Compute the prediction probability E_{st} from $\prod_{k \in R_s, k \neq t} E_{ks} Q_s E_{us}$, where E_{ks} is the prediction probabilities computed and stored in node s during the upward propagation, E_{us} is the prediction probability from node u to s, which is the parent node of s, Q_s is the posterior probability of s.

In order to widely spread the beliefs, during the downward propagation, one can propagate the belief to more nodes at finer scales. This can help to smooth the disparity field even when there are large regions where the measurements are noisy.

The propagation of information using above scheme is non-iterative, which is a clear advantage over the MRF model. If the original image size is N pixels, the maximum number of scales is $\log N$. The number of propagations between scales is proportional to N. Therefore the total run time of the propagation scheme is $O(N \log N)$, a significant improvement over the MRF model.

5.5 Experimental Results

We implemented the multi-scale propagation algorithm using the same Gaussian pyramid in the direct pooling method. The complete algorithm is summarized in the following:

1. Obtain the binocular measurements at every scale and orientation.

2. Compute the likelihood function and joint likelihood functions using Eq. (4.6) and Eq. (4.9).

3. Upward propagation and downward propagation: use the propagation scheme summarized at the end of Section 5.3 to compute the posterior probabilities. The propagation is performed separately at each orientation. The final posterior probabilities are obtained by multiplication of the posterior probabilities from all three orientations.

4. Use MAP to get the disparity map.

The algorithm has been tested on several real and synthetic image pairs. The disparity map is obtained using MAP from the posterior. All disparity maps have half pixel resolution. To achieve higher sub-pixel resolution, one can use the method discussed in Section 5.1. For the purpose of comparison, we show the disparity maps of the SRI tree sequence using a quadratic transition function in Fig. 5.14 and a square root transition function in Fig. 5.15. The quadratic model is given by Eq. (5.3) and Eq.(5.4), while the square root model is given by Eq. (5.5) and Eq.(5.6). The quadratic model has better overall quality than the square root model, due to the fact that the quadratic model is a good model to smooth the disparity field, even though it may work poorly at object boundaries. The square root model may be a better model at the object boundary, at the expense of a noisier disparity field. However, here we deal with multi-modal distributions, where multiple disparities have high probability at a point. Since we propagate the evidence for all these probabilities, it may be fine to have a simple quadratic (Gaussian) distribution for the neighborhood interaction. The probability density functions of both models are shown in Fig. 5.11. The value of B in Eq. (5.3 - 5.6) should be chosen such that the detail of the scene is kept while the errors are smoothed out. If B is too large, it may lead to the loss of detail. While if B is too small, it may not be able to smooth out the errors. In practice, we found

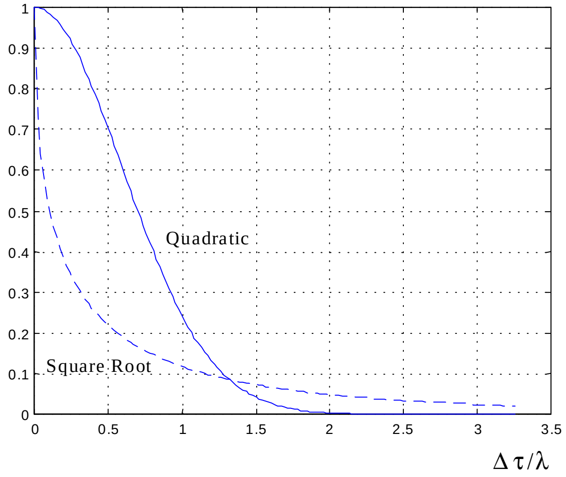

Figure 5.11: The probability density functions for the square root model and the quadratic model. The quadratic model is given by Eq. (5.3) and Eq.(5.4), while the square root model is given by Eq. (5.5) and Eq.(5.6). We choose $B = 15$ for the quadratic model and $B = 1$ for the square root model.

that $B = 15$ for the quadratic model and $B = 1$ for the square root model achieve best balance.

Fig. 5.16 shows the disparity map for the SRI tree with information spreading to more nodes at the finer scale during the downward propagation. One may note that there are fewer large regions of errors, due to the smoothing effect of more widely spread information.

Fig. 5.13 is the disparity map of the lamp/head sequence

The disparity estimates from the pentagon image pair and the lamp/head sequence are shown in Fig. 5.12 and Fig. 5.13, respectively. They are estimated using the quadratic model and a widely spread downward propagation. Notice that the disparity map recovers more details of the depth scene than the local weighted phase

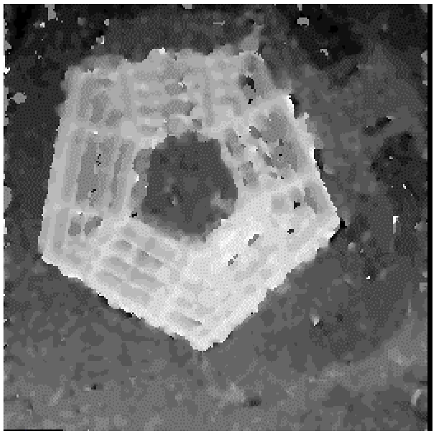

Figure 5.12: Disparity estimate of the Pentagon image pair using the multi-scale Stochastic model. The transition function is a quadratic model with a widely spread downward propagation. The disparity map has half pixel resolution.

correlation method, but without making the estimates noisier. The result is compatible with the direct pooling method.

The multi-scale model is our initial attempt to combine information. Currently it does not have explicit occlusion detection and handling. However, the SRI tree sequence, as well as many other stereo pairs contain occluded regions, which will inevitably cause incorrect matching for our algorithm. One can see that there are several obvious false matches in some region near the trunk of the tree, due to occlusion. One of the advantages of the new algorithm is that it may be possible to incorporate an occlusion detection mechanism into the model. [Bel95, GLY95] provide some work on the detection of occlusion within Bayesian framework, which may help us improve the algorithm.

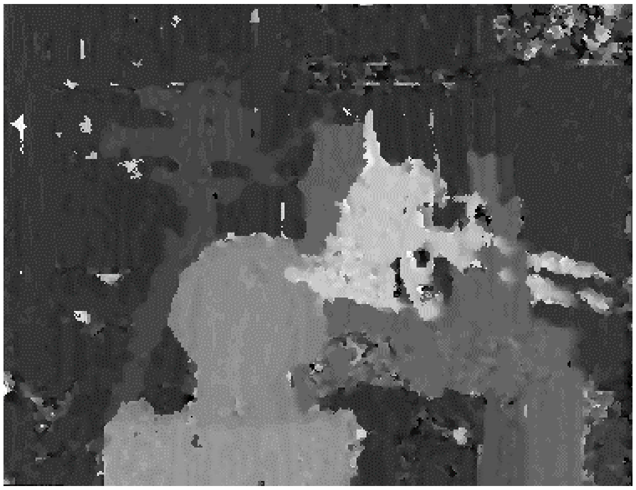

Figure 5.13: Disparity estimate of the lamp/head sequence using the multi-scale Stochastic model. The transition function is a quadratic model with a widely spread downward propagation. The disparity map has half pixel resolution.

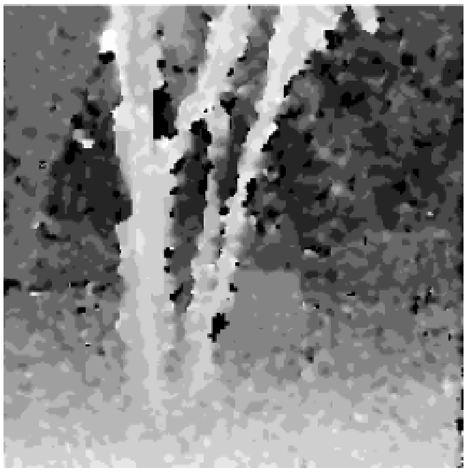

Figure 5.14: Disparity estimate of the SRI tree sequence (frame 2 and frame 4) using the multi-scale Stochastic model. The transition function is a quadratic model. The disparity map has half pixel resolution.

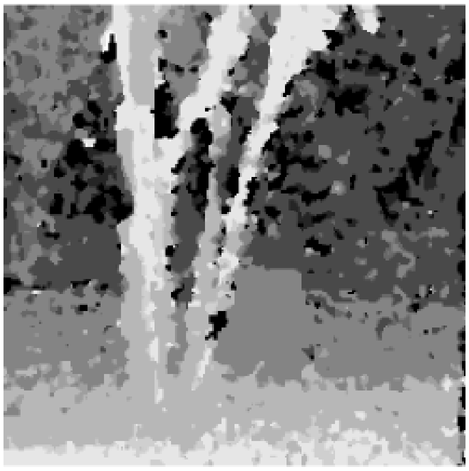

Figure 5.15: Disparity estimate of the SRI tree sequence (frame 2 and frame 4) using the multi-scale Stochastic model. The transition function is a square root model. The disparity map has half pixel resolution.

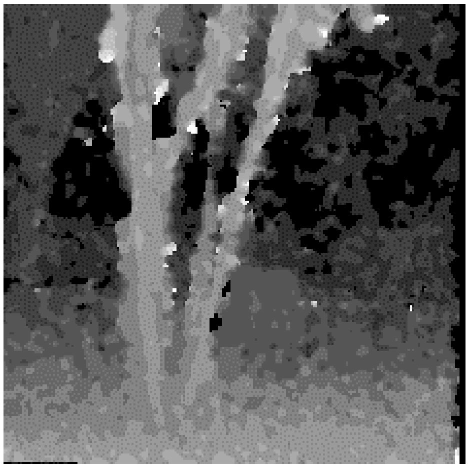

Figure 5.16: Disparity estimate of the SRI tree sequence (frame 2 and frame 4) using the multi-scale Stochastic model. The transition function is a quadratic model with a widely spread downward propagation. The disparity map has half pixel resolution.

5.6 Summary

In this chapter, we investigated two ways to combine information from different channels. The simple way is to take the product of the likelihood functions across scales and orientations by assuming the independence of measurements and a uniform prior model. This approach is similar to the local weighted correlation method by Fleet [Fle94].

A more sophisticated way is to exploit the spatial coherence of the measurements by using a multi-scale Markov prior model. This results in a Bayesian network without loops. A propagation scheme is designed to avoid the double counting of information, which typically happens in Bayesian network with loops. The resulting algorithm is non-iterative and does not suffer from the problem associated with coarse-to-fine control strategies.

Chapter 6

Conclusions and Future Work

As discussed in chapter 2, Phase-based methods for stereo matching have many advantages over other methods. Neurophysiological research has also shown that phase-based measurements comprise the first stage of disparity processing in the primary visual cortex of many mammals and in the visual wulst of the owl. Therefore, phase-based methods are not only important for computer vision research, they also play important role in modeling biological stereopsis. Despite the success of phase-based method in modeling the early stage of disparity processing, the subsequent stages that combine the measurements to infer a unique disparity map are unknown. The commonly used coarse-to-fine control strategy in computer vision may not be suitable for modelling this process. That is because, with the coarse-to-fine approach, a poor estimate at the coarse scale leads to incorrect estimate at the fine scale, from which the algorithm cannot recover. There is also evidence against the use of coarse-to-fine control strategy in biological vision [MDA94].

In this thesis, we formulate phase-based disparity estimate with Bayesian approach. We use Bayesian framework to determine the optimal way to combine the results of phase-based method in different channels. This can potentially be useful to model the second stage of disparity processing in primary visual cortex. Current computer vision techniques combine these estimates in a somewhat ad hoc way, assuming that left and right images are simple translation of one another. However, this assumption is valid only when the 3-D surfaces are frontopararell. The Bayesian approach is more flexible because it can be applied to a broad range of surfaces. This is achieved through the use of different prior models for different kinds of surfaces. In order to incorporate the phase-based method into a probabilistic framework, we provide the Phase-based method a probabilistic basis.

6.1 Conclusions

The phase-based measurement used in this thesis in the binocular measurement. The main contribution of this thesis is the development of a likelihood function for the binocular measurement for use in a Bayesian framework. The related works we have done include:

- Identification and modeling the sources of variability in the binocular measurements.

- Empirical derivation of the form of the likelihood function for the binocular measurements at different scales and orientations.

- Fitting the form of the likelihood function by a Beta distribution, which is invariant in scale but different across orientation.

- Formulation of a joint likelihood function for measurements at different pre-shifts for a single scale and orientation.

We have also completed some initial work related to the combination of measurements across scales and orientations:

- Implementation of the simplest way to combine the measurement over scales and orientations, by taking the product of the joint likelihood function across scales and orientations. By doing so, we assume the independence of the measurements at different scales and orientations, and the uniform prior over disparity maps. Compare this implementation to the local weighted phased-correlation method by Fleet [Fle94].
- Development of an algorithm based on a multi-scale Markov prior model that prefers smooth disparity fields and small disparities. Design and implementation of an information propagation scheme for the multi-scale model that avoids the "double counting" of information. Implementation and testing of the above algorithm and the local weighted phase-correlation algorithm.

We developed the new algorithm in order to show the feasibility of using the Bayesian approach with phase-based measurements. Therefore the performance is not our major concern. However, the new algorithm has shown many potential advantages over existing MRF-based approaches. Currently, many MRF-based algorithms incorporate smoothness models that require iterative procedures, with coarse-to-fine propagation of estimates. They are usually slow to converge and are therefore not suitable for real-time applications. The multi-scale algorithm here does not assume coarse-to-fine and it is computed in fixed time in terms of the number of pixels.

6.2 Future Work

The sources of variability in the binocular measurement include noise, smooth but non-constant surface, variation in the instantaneous frequency, discontinuities at object boundary, and deformation/scale change. We only model the first three sources. How to explicitly model of discontinuities at object boundary, and Deformation/Scale change remains for future research.

Currently, when we derive the joint likelihood function, we first assume the independence of the measurements at different pre-shift and take the product of the likelihood functions to form the joint likelihood function. However, since the measurements are, in fact, correlated, the joint likelihood function derived in this way has sharper peaks than they should have. The simple approach to overcome this problem is to raise the joint likelihood function to the power of some value. We may want a more elegant way to deal with the correlation.

The multi-scale model does not explicitly take the discontinuities and occlusion into account. There is work on how to detect and handle the discontinuities and occlusion within the Bayesian framework with MRF as prior model [Bel95, GLY95]. There is also similar work on detection and tracking of motion discontinuities [BJ99]. Psychophysical evidence [GLY95, GB88] also indicates that the human visual system takes advantage of the occluded regions for obtaining depth information. However, how to incorporate occlusion detection mechanism in the multi-scale prior model needs further research. We may also need a better propagation function that supports the piece-wise smoothness and preserves the object boundary and depth discontinuities.

Bibliography

[Bar89] S. T. Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3:17–32, 1989.

[Bel95] P. N. Belhumeur. A Baysian approach to binocular stereopsis. *International Journal of Computer Vision*, 19:237–260, 1995.

[BJ99] M. J. Black and Fleet D. J. Probabilistic detection and tracking of motion discontinuities (accepted october, 1999). *International Journal of Computer Vision*, 1999.

[BS94] C. A. Bouman and M. Shapiro. A multiscale random field model for Baysian image segmentation. *IEEE Trans. On Image Processing*, 3(2):162–177, 1994.

[Cas96] K. R. Castleman. *Digital image processing.* Prentice Hall, 1996.

[CC85] R. Chellapa and S. Chatterjee. Classification of textures using gaussian markov random fields. *IEEE Transactions on Speech and Signal Processing*, 33:959–963, 1985.

[DOF91] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature*, 352:156–159, 1991.

[FA91] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. PAMI*, 13:891–906, 1991.

[FJ90] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5:77–104, 1990.

[FJ93] D. J. Fleet and A. D. Jepson. Stability of phase information. *IEEE Trans. PAMI*, 15(12):1253–1268, 1993.

[FJJ91] D. J. Fleet, A. D. Jepson, and M. Jepson. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.

[Fle94] D. J. Fleet. Disparity from Local Weighted Phased-Correlation. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 48–56, 1994.

[FWH96] D. J. Fleet, H. Wagner, and D. J. Heeger. Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research*, 36(12):1839–1857, 1996.

[GB88] B. Gillam and E. Borsting. The role of monocular regions in stereoscopic displays. *Perception*, 17:603–608, 1988.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6(6):721–741, 1984.

[GG91] D. Geiger and F. Girosi. Parallel and deterministic algorithms for mrfs: surface reconstruction. *IEEE Trans. PAMI*, 13(5):401–412, 1991.

[GLY95] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[IB96] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conf. Computer Vision*, pages 343–356, 1996.

[JJ89] A. Jepson and M. Jenkin. The fast computation of disparity from phase difference. In *IEEE CVPR*, pages 398–403, 1989.

[JJ94] M. Jenkin and A. Jepson. Recovering local surface structure through local phase difference measurements. *CVGIP: Image Understanding*, 59:72–93, 1994.

[KH75] C. D. Kuglin and D. C. Hines. The phase correlation image alignment methods. In *Proceedings IEEE Conf. On Cybernetics and Society*, page 163, 1975.

[KO93] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. In *Proc. 1991 IEEE International Conference on Robotics and Automation*, pages 1088–1095, 1993.

[KWS75] T. J. Keating, P. R. Wolf, and F. L. Scarpace. An improved method of digital image correlation. *Photogrammetric Engineering and Remote Sensing*, 41:993–1002, 1975.

[LB95] A. Luo and H. Burkhardt. An intensity-based cooperative bidirectional stereo matching with simultaneous detection of discontinuity and occulsion. *International Journal of Computer Vision*, 15:171–188, 1995.

[LD93] S. Lakshmanan and H. Derin. Gaussian markov random fields at multiple resolution. *Markov Random Fields, Theory and Applications*, pages 131–157, 1993.

[LKW94] M. R. Luettgen, W. Clem Karl, and A. S. Willsky. Efficient multicsale regularization with application to the computation of optical flow. *IEEE Transactions on image processing*, 3(1):41–64, 1994.

[LKWT93] M. Luettgen, W. Karl, A. Willsky, and R. Tenny. Multiscale represnetation of Markov Random Fields. *IEEE Trans. Signal Processing*, 41(12):3377–3396, 1993.

[MDA94] H. Mallot, S. Dartsch, and P. Arndt. Is correspondence search in human stereo vision a coarse-to-fine process? *Technical Memo No. 4, Max Plank Institute for Biological Cybernetic, Tubingen, Germany*, 1994.

[MMP87] J. Marroquin, S. Mitter, and T. Poggio. Probability solution of ill-posed problems in computational vision. *Journal of American Stat. Assoc.*, 82:76–89, 1987.

[MP77] D. Marr and T. Poggio. *A theory of human stereo vision.* A.I. Memo451, Artificial Intelligence Laboratory, MIT, 1977.

[Muk96] P. Mukhopadhyay. *Mathematical statistics.* New Central Book Agency, India, 1996.

[Nal93] V. S. Nalwa. *A guided tour of computer vision.* Addison Welsley, 1993.

[Nis84] H. K. Nishihara. Practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545, 1984.

[ODF90] I. Ohzawa, G. DeAngelis, and R. Freeman. Stereoscopic depth discrimination in the visual cortex: Neuron ideally suited as disparity detector. *Science*, 249:1037–1041, 1990.

[Pea88] J. Pearl. *Probability reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann, 1988.

[PTK85] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[San88] T. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418, 1988.

[Sto86] J. Stoke. Image matching with phase shift methods. In *ISPRS Comm III. Symposium*, pages 638–652, 1986.

[Wei97] Y. Weiss. Interpreting images by propagating bayesian beliefs. In *Advances in Neural Information Processing Systems*, pages 908–915. 1997.

[Wei99] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation (To appear)*, 1999.

[Wen94] J. Weng. Image matching using windowed fourier phase. *International Journal of Computer Vision*, 12:211–236, 1994.

[WF93] H. Wagner and B. J. Frost. Disparity-sensitive cells in the owl have a characteristic disparity. *Nature*, 364:756–798, 1993.

Appendix A

A.1 Steerable Filters

Oriented filters have found extensive use in many computer vision and image-processing task, such as edge detection, texture analysis, image compression and motion analysis, etc. We also use oriented filters in our algorithm of stereo matching for the reasons given in section 2.5.2. In many tasks, it is useful to be able to tune the orientation of the filters to arbitrary orientation. It is a tedious work to design filters of all orientations and it takes a lot of space to store the kernels of the filters. One natural question arises: is it possible to design a set of filters of different orientation and use them as basis functions to synthesize filters with arbitrary orientation?

It has been proved possible to do so with the concept of "steerable filters" introduced in [FA91]. The term "steerable filters" is used to describe a class of filters in which a filter of arbitrary orientation is synthesized as linear combination of a set of "basis filters"(Fig. A.1). Formally

$$f^{\theta}(x,y) = \sum_{j=1}^{M} k_j(\theta) f^{\theta_j}(x,y) \tag{A.1}$$

or in polar coordinates

$$f^{\theta}(r,\Phi) = \sum_{j=1}^{M} k_j(\theta) g_j(r,\Phi) \tag{A.2}$$

where $r = \sqrt{x^2 + y^2}$ and $\Phi = \arg(x,y)$. θ is the orientation the filter is tuned to.

To design a steerable filter, besides the design of the basis filters, we also need to know the minimum number of basis filters that are sufficient for steering and the coefficients for the basis filters. As an example, we show the design of a steerable quadrature filter pair based on the frequency response of the second derivative of a Gaussian (G2). This is the steerable filter we use in our stereo matching algorithm.

The second derivative of a Gaussian is $f(x,y) = G_2^{0^\circ} = (4x^2 - 2)e^{-(x^2+y^2)}$. The superscript 0° indicates that its orientation is 0°. As shown in [FA91], three basis filters are sufficient for steering. The orientations of the basis filters are 0°,60° and 120°. The coefficients are defined by

$$k_j(\theta) = \frac{1}{3}[1 + 2\cos(2(\theta - \theta_j))] \tag{A.3}$$

Thus a G_2 filter with orientation θ can be obtained by

$$G_2^{\theta} = k_1(\theta)G_2^{0^\circ} + k_2(\theta)G_2^{60^\circ} + k_3(\theta)G_2^{120^\circ} \tag{A.4}$$

The Hilbert transform of G_2, i.e. H_2, can be approximated as a polynomial times a Gaussian. One can use a least squares fit to find the polynomial. Given in [FA91], the approximation of H_2 is

$$H_2 = (-2.205x + 0.9780x^3)e^{-(x^2+y^2)} \tag{A.5}$$

The same technique is used to find the basis filters and the coefficients to steer H_2. Since H_2 has higher order than G_2, it takes four basis filters as interpolation functions.

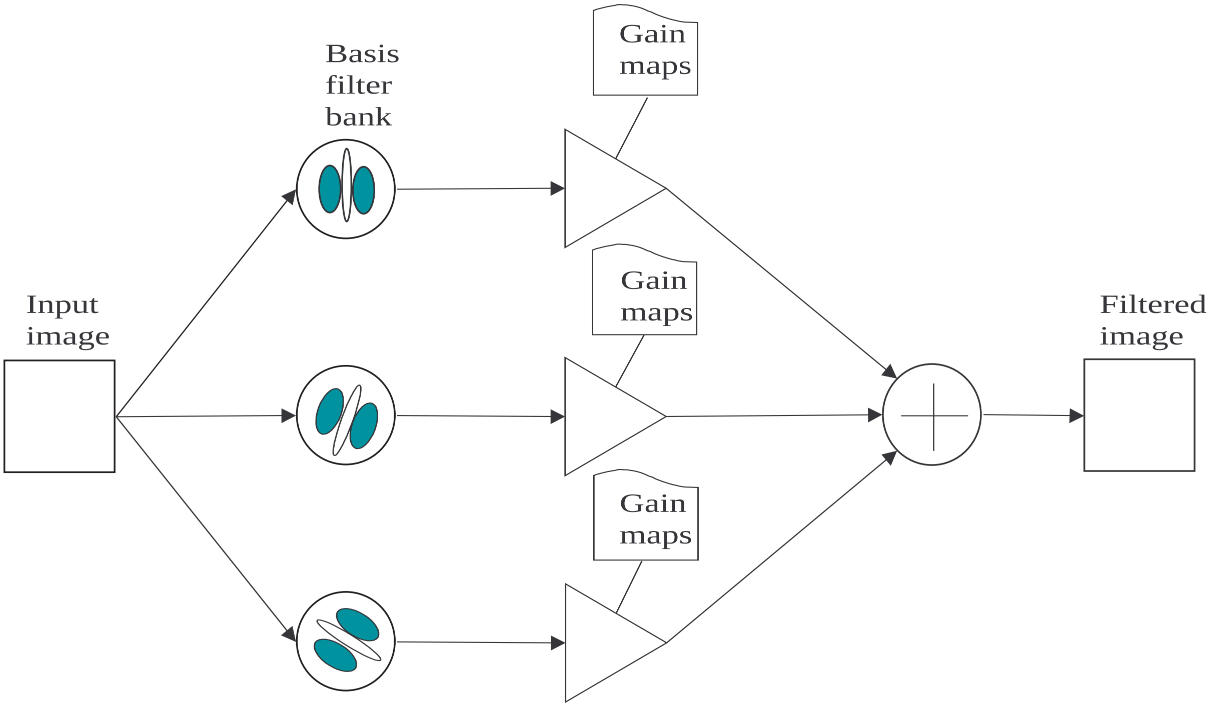

Figure A.1: Steerable Filter Architecture. A bank of basis filters first processes the input image. The outputs are then multiplied by a set of gain maps followed by a summation. This results in an output image that is the equivalent of one filtered by a certain orientation.

Note that all the basis filters for G_2 and H_2 are in the form of a polynomial in x and y times a Gaussian, thus they are x-y separable. For example, $G_2^{60^\circ}$ can be written as

$$G_2^{60^\circ} = 1.843xye^{-(x^2+y^2)} = 1.843xe^{-x^2} \cdot ye^{-y^2} \tag{A.6}$$

Thus one can implement $G_2^{60^\circ}$ by first applying the one-dimensional filter $1.843xe^{-x^2}$ horizontally to the input image and then ye^{-y^2} vertically. The x-y separable property of these filters significantly reduces the computational cost.

Appendix B

B.1 Stochastic Relaxation

It is well known that the global optimization problem posed by the MRF model is non-trivial. If the image size is $N \times M$, and the disparity takes the value from $-d_{max}$ to d_{max}, then the Bayesian approach with a MRF prior model requires us to find a disparity map from $(2d_{max})^{N \times M}$ candidate maps that minimizes the energy function E_D. To obtain the optimal solution by an exhaustive search method, the computational complexity is an exponential $O((2d_{max})^{N \times M})$, which in most cases is computationally prohibitive. One has to turn to sub-optimal solutions, such as *stochastic relaxation* [GG84], which is often described as *simulated annealing* due to its conceptual similarity to a physical process called annealing.

Simulated annealing is an iterative algorithm widely used in many applications involving combinatorial optimization, including the famous NP complete problems such as travelling salesperson problem, Hamilton circuits, etc. For convenience, we describe this algorithm here in the context of stereo matching. Let the function to be minimized be $E(D)$, where D is the disparity map. The algorithm can be described as follows:

1. Begin with a random disparity map D_0 and an initial temperature parameter $T = T_0$.

2. At step k, perturb D_k by $\hat{D}_{k+1} = D_k + \Delta D$ and compute $\Delta E = E(\hat{D}_{k+1}) - E(D_k)$.

3. if $\Delta E < 0$, accept the change, that is, $D_{k+1} = D_k + \Delta D$. If $\Delta E > 0$, accept the change only with probability $p = e^{-\Delta E/T}$.

4. Decrease T according to some temperature schedule.

5. If the energy becomes stable and the temperature is very low, then stop; otherwise go to step 2.

The perturbation function can be realized in many ways. For example, disparity can be randomly increase or decrease by one, or an entirely new disparity map can be randomly chosen from some fixed range. A number of temperature schedules are available. The one used by Geman and Geman [GG84] is

$$T(k) = \frac{C}{log(1+k)}, \quad 1 \leq k \leq K \tag{B.1}$$

where $T(k)$ is the temperature during the kth iteration, K is the total number of iterations. In general, a more slow cooling schedule is more likely to attain a final state close to a global optimum.

One can perform simulated annealing directly on a pair of stereo images, but the convergence would be rather slow if the images have large size and the disparity range is large. In recent years, many algorithms have been developed to speed up the computation of MRF. A more efficient method is to use the coarse-to-fine strategy. One obtains a series of images with different resolution by blurring and sub-sampling the original images. At a coarse level where the image has low resolution, the size of image and the range of disparity are small. Therefore simulated annealing can achieve fast convergence. The result obtained at the coarse level can be used as the starting point for the next finer level to speed up the convergence at that level. Note that by doing so, we still use the iterative algorithm on each scale, only with improved initial estimation to speed up the convergence. In recent years, researchers found that the MRF model can be replaced by a multi-scale stochastic model, which can eliminate the iterative procedures while achieve compatible or even better results, with much less computation [BS94, LKW94]. [LKWT93] has developed a theoretical framework to justify the model.

Appendix C

C.1 Parameter Estimate of Beta Distribution

A random variable X is said to have a Beta distribution with parameters a, b ($a > 0, b > 0$), if its probability distribution function is given by

$$f(x; a, b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1, a > 0, b > 0 \\ 0, & \text{otherwise} \end{cases} \tag{C.1}$$

where $B(a, b)$ is the Beta function

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt, \; (a, b > 0) \tag{C.2}$$

Because $C(x_0, \tau_0 \pm \Delta\tau)$ ranges from -1 to 1, Eqn. (C.1) can be rewritten as

$$f(x; a, b) = \begin{cases} \frac{1}{2B(a,b)} (\frac{x+1}{2})^{a-1}(1 - \frac{x+1}{2})^{b-1}, & -1 < x < 1, a > 0, b > 0 \\ 0, & \text{otherwise} \end{cases} \tag{C.3}$$

To estimate the parameters of the Beta distribution, one can use the method of moments (MM) to have a raw estimate, then use the raw estimate as an initial value to iteratively solve a set of equations derived by the Maximum Likelihood Estimate (MLE) method to get more a accurate estimate.

The first and second moment of the Beta distribution is

$$\mu_1 = \mathrm{E}(X) = \frac{a}{a+b}, \qquad \mu_2 = \mathrm{E}(X^2) = \frac{a(a+1)}{(a+b)(a+b+1)} \tag{C.4}$$

For a Beta distribution in the form of Eq. (C.3) the first and second moment become

$$\mu_1' = 2\mu_1 - 1 = \frac{a-b}{a+b}, \qquad \mu_2' = 4\mu_2 - 4\mu_1 + 1 = \frac{(a-b)^2 + (a+b)}{(a+b)(a+b+1)} \tag{C.5}$$

Solving the above equations, we obtain

$$a = \frac{(\mu_1'\mu_2' - \mu_1' + \mu_2' - 1)}{2(\mu_1'^2 - \mu_2')}, \qquad b = \frac{(-\mu_1'\mu_2' + \mu_1' + \mu_2' - 1)}{2(\mu_1'^2 - \mu_2')} \tag{C.6}$$

where $\mu_1' = \mathrm{E}(X')$ and $\mu_2' = \mathrm{E}(X'^2)$, X' is a random variable that has the distribution given in Eq. (C.3).

Usually the parameters estimated by the method of moment are not very accurate. To obtain a more accurate estimate, we can use the Maximum Likelihood Estimate (MLE). Given n samples $x_1, x_2, ...x_n$, the MLE of a and b of a Beta distribution is:

$$\begin{aligned} \frac{-dB(a,b)}{db}\frac{n}{B(a,b)} + \sum_{i=1...n}^{n} \ln((1-x_i)/2) = 0 \\ \frac{-dB(a,b)}{da}\frac{n}{B(a,b)} + \sum_{i=1...n}^{n} \ln((1+x_i)/2) = 0 \end{aligned} \tag{C.7}$$

where $B(a,b)$ is the Beta function. The above nonlinear equations can not be solved in closed form; an iterative method for finding the roots has to be employed. We could use the initial value obtained by the method of moment to start the iterative procedure. In our experiment, we found that the estimates by the method of moment along are almost as good as the optimal estimates found by the iterative ML method. Therefore, we use method of moment to estimate the parameters.

Vita

Experiences **Assistant Engineer**
1994 - 1996
Engineering Department
Winger Electronics Corporation
China

Education **Master of Science**
1998 - 2000
Dept. Of Computing and Information Science
Queen's Unviversity

Bachelor of Electrical Engineering
1989 -1993
Dept. Of Electrical Engineering
Zhejiang University
China