ABSTRACT

Title of Dissertation: VIDEO LINKING

Robert Pless, Doctor of Philosophy, 2000

Dissertation directed by: Professor Yiannis Aloimonos
Department of Computer Science

When a video is captured by a camera moving through a scene, the changing images encode the camera motion parameters and the structure of the scene. This proposal studies models of scene structure, and their relationships to simple functions of the spatial and temporal derivatives in the sequence of images. This study has three major thrusts. First, we study noise properties of image measurements in order to determine what measurements are robust. Second, we provide and analyze an algorithm to simultaneously provide a camera motion estimate and a piecewise planar model of the scene which respects scene depth discontinuities. Third, we study methods of matching and reconciling information from disparate camera viewpoints. These goals all directly relate to the problem of creating viewpoint independent descriptions of arbitrary scenes from

general camera motions. This relates to the study of human perception and has applications to Virtual Reality, Video Editing, Graphics, and Robotics.

VIDEO LINKING

by

Robert Pless

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2000

Advisory Committee:

Professor Yiannis Aloimonos, Chairman/Advisor
Professor Azriel Rosenfeld
Professor Larry Davis
Associate Professor Samir Khuller
Assistant Professor David Poeppel

Dedication

To my family.

Acknowledgements

This work would not have been possible without the help of many kind, thoughtful, and considerate people. Yiannis Aloimonos, as my advisor, provided guidance, support, and inspiration throughout my studies at Maryland. Other members of our research group, Cornelia Fermüller, David Shulman, LoongFah Cheong, Gregory Baratoff, Tomas Brodsky, Brad Stuart, and Jan Neumann, have all been unceasingly helpful and encouraging. Jordan Landes and Matthew Katsouros contributed greatly to my positive impression of the Department of Computer Science at the University of Maryland. Special thanks go to Patrick Baker, my friend and now my colleague of many years, who is eternally willing to share new insights and new ideas about vision, science, and the changing political structures in the world today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A video is a sequence of images taken by a camera moving through a scene. The act of taking a picture defines a relationship between three things: the scene, the position and calibration of the camera, and the image. This dissertation is concerned with creating accurate models of the scene from these images. All approaches to this problem can be characterized in terms of the representational assumptions that they make – or, how they answer the following specific questions:

- What measurements can be made on one or multiple images?
- What assumptions are made about the camera projection and calibration?
- What are allowable camera motions?
- What are appropriate scene models?

I choose to consider a system with the following attributes. The scene model is a set of small planar patches of arbitrary orientation. The camera motion is arbitrary but continuous. I assume the camera is a pinhole camera, and that I have complete knowledge of the camera calibration (the focal length, etc.). I use only spatio-temporal derivatives of the image intensity, as opposed to matching sets of feature points.

These choices are not arbitrary; rather, they define a vision system appropriate to analyze data from standard video cameras in a large variety of natural scenes. The choice to use well segmented planar patches of arbitrary orientation reflects a trade off between the representational complexity and accuracy of the scene model. While the optics of most video cameras are quite complicated, a pinhole camera model strikes a balance between the accuracy and complexity of the projection model. The assumption of perfect knowledge of the camera calibration serves to focus this study; recent work on self calibration from video sequences makes this a plausible assumption. Continuous camera motion is appropriate for a hand held video camera that is continuously capturing images, any further limitations on the camera movement would detract from the general usefulness of the algorithm.

This dissertation has three major contributions. The first is an explanation of why image derivatives are robust measurements of changing images. The alternative, which is most commonly used in computer vision algorithms, is to

compute the optic flow, a two dimensional vector field describing the motion of points from one image to the next. An analysis of common techniques to find this optic flow field is presented in Chapter 3. In the presence of noise in the image, there is a bias in the computed optic flow field. This bias is dependent upon the local image motion, the local image texture, and the amount of noise in the system. This formulation accurately predicts several classes of perceptual illusions, involving segmentation of moving patterns and the mis–estimation of motion.

The second contribution includes an implementation of a method to compute the instantaneous velocity of a video camera and the scene depth. This implementation includes techniques for automatically finding an appropriate segmentation of the scene. The algorithm then essentially calculates the value of an error function over the space of possible translational directions of the camera. Chapter 4 gives an experimental study of the topography of this error surface. The main result is that in many cases the minimum of this function is ambiguous — and therefore it is impossible to accurately solve for the direction the camera is moving. It is possible to consistently find a small set of plausible translation directions. This set tends to be extended along a line, these are translation directions that lie along a "valley" or extended minima of the error function.

The final contribution of this dissertation is a study of how to use visual information captured from multiple viewpoints. This additional information serves

to fix the camera motion estimate for each viewpoint along its valley, and defines a method for combining scene reconstructions. Chapter 5 defines the geometry involved in the linking process, both in the simple case of linking subsequent frames from video sequence, and the more powerful constraint of linking camera viewpoints that are further apart. This process involves the matching of 3D representations generated from each viewpoint. Because it uses more data, it is intrinsically more powerful than current state of the art image matching techniques.

Chapter 2

Preliminaries

There are a large number of events that have to be modeled in a video analysis system. The camera moves along some path through space. As it takes each image, the location of the camera and its internal parameters define the precise relationship between the position of a point in the space, and the position of the image of that point. The images change continuously as the camera moves. Measuring how these images change defines a constraint between the motion of the camera and the scene. These constraints can be combined to estimate the scene structure and camera motion on the basis of these image measurements. This chapter details the mathematics that can specify these events, and the previous work that has been done along similar paths.

2.1 How an Image is Formed

Since the initial input is a sequence of images captured by a camera, it is convenient to work with a Cartesian coordinate system $OXYZ$ attached to the camera. Point O is the nodal point of the camera and the Z axis is aligned with the optical axis. How does the 3D world project onto the imaging surface (retina)? Even an off-the-shelf video camera is a quite complicated optical instrument whose precise modeling may require a lot of effort. We need to strike a balance between simplicity of the model and its accuracy. The choice most often made in Computer Vision literature, a simple pinhole camera model, is sufficient in most cases. The images are formed by perspective projection on the retina. If the ignored optical effects, such as radial distortion, are too strong, the model can still be used if a pre-calibration step is performed.

The planar retina (also called the image plane) is illustrated in Figure 2.1. The image plane is perpendicular to the optical axis of the camera at distance f from the nodal point. The most simple case is a normalized camera, where the exact mapping between 3D directions and image points is known, and the focal length $f = 1$.

With perspective projection, the image of a scene point $\mathbf{R} = (X, Y, Z)$ is:

$$\mathbf{r} = f \frac{\mathbf{R}}{\mathbf{R} \cdot \hat{\mathbf{z}}} \tag{2.1}$$

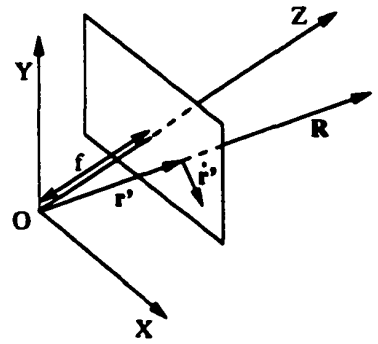


Figure 2.1: Image formation on a planar retina.

where $\hat{\mathbf{z}} = [0, 0, 1]^{\mathrm{T}}$ is the unit vector in the direction of the optical axis. The depth of the scene point, Z, is the dot product $\mathbf{R} \cdot \hat{\mathbf{z}}$, the distance along the optical axis from the camera center to the scene point.

In an uncalibrated camera, the relationship between the position of points in the world and the position of their projections on the image is dependent upon the calibration matrix K. The mapping between a scene point $\mathbf{R}$ and the corresponding image point $\mathbf{r}$ can be concisely written as [14]

$$\mathbf{r} = \frac{\mathbf{KR}}{\mathbf{R} \cdot \hat{\mathbf{z}}} \tag{2.2}$$

where $\hat{\mathbf{z}}$ is again the unit vector in the direction of the Z axis. Knowledge of this matrix $\mathbf{K}$ allows an image to be warped to appear as if it were taken by a normalized camera.

The problem of determining $\mathbf{K}$ from arbitrary images is termed self-calibration, early studies demonstrated the theoretical feasibility of self calibration [43, 14, 23]. More recent work gives a direct algorithm to find the parameters of the calibra-

tion matrix, using only image derivatives from a video sequence of at least four images which include at least two different rotational motions [9]. The work in this thesis assumes that the camera has been calibrated with this technique, or an equivalent one, so the projection of world points onto an image can be modeled with the simpler form of equation 2.1.

2.2 Image Measurements

For a dense image sequence, the images captured by a camera are a dense sampling of the image intensity function I. $I(x, y, t)$ is the image intensity at image position (x, y) at time t. This sampling of the intensity function is smoothed with a Gaussian filter to allow the use of first order derivatives. In this study, "image measurements" refers to the spatial and temporal derivatives of the Gaussian smoothed images, and these derivatives I_x, I_y, I_t are assumed to be derivatives of the smoothed images.

The relative motion between the camera and the scene induces a 3D velocity field with respect to a coordinate system fixed at the camera center. The projection of this 3D motion onto the image plane is called the motion field. The observed motion of brightness patterns across the image is the optic flow field. Under some conditions, such as a translating Lambertian surface with constant illumination [62], these fields are identical. In extreme cases, they may have little

relationship to each other [29], but typically, in regions of some variance of image intensity, the optic flow field is a close approximation to the motion field.

The standard assumption allowing visual motion analysis is that the same point in space always has the same intensity in the image. This image brightness constancy assumption leads to the constraint [29]:

$$0 = \frac{dI}{dt} = \frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} \tag{2.3}$$

This equation relates the image velocity $\mathbf{u} = (\frac{dx}{dt}, \frac{dy}{dt})$ with the spatio-temporal derivatives of the image intensity function.

2.3 Camera Motion

The instantaneous motion of a camera relative to the scene can be described as a translational velocity $\mathbf{t}$ and rotational velocity $\boldsymbol{\omega}$ around the nodal point of the camera. One can then derive the motion of a point on the image plane to be [41]:

$$\mathbf{u}(\mathbf{r}) = \frac{1}{\mathbf{R}\cdot\hat{\mathbf{z}}}\mathbf{u}_{tr}(\mathbf{r}) + \mathbf{u}_{rot}(\mathbf{r}) = -\frac{1}{Z}(\hat{\mathbf{z}} \times (\mathbf{t} \times \mathbf{r})) + \frac{1}{f}(\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r}))) \tag{2.4}$$

a vector which has a component dependent upon both the translational velocity and the depth of the point, and a component solely related to the rotation of the camera.

From image measurements alone, it is not possible to determine the vector $\mathbf{u}(\mathbf{r})$, because equation 2.3 gives only one constraint on this two-dimensional

vector. Only the component of this vector in the direction of the image gradient is constrained. The constrained motion in the direction of the image gradient is called the normal flow. If $\mathbf{n}$ is a unit vector in the gradient direction on the image plane (so that vector $\mathbf{n} \cdot \hat{\mathbf{z}} = 0$), the normal flow at image point $\mathbf{r}$ is:

$$u_n(\mathbf{r}, \mathbf{n}) = -\frac{1}{Z}(\hat{\mathbf{z}} \times (\mathbf{t} \times \mathbf{r})) \cdot \mathbf{n} + (\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r}))) \cdot \mathbf{n} \tag{2.5}$$

This normal flow vector is considered to be a direct image measurement because it is a function of the image derivatives at $\mathbf{r}$ [29]:

$$u_n(\mathbf{r}, \mathbf{n}) = \left\langle \frac{I_x I_t}{\sqrt{I_x^2 + I_y^2}}, \frac{I_y I_t}{\sqrt{I_x^2 + I_y^2}} \right\rangle \tag{2.6}$$

For every pixel i, we collect the measurable image quantities to define:

$$\vec{\mathbf{a}}_i = \begin{pmatrix} -I_{x_i} \\ -I_{y_i} \\ I_{x_i} x + I_{y_i} y \end{pmatrix} \text{ and, } \vec{\mathbf{b}}_i = \begin{pmatrix} I_{x_i} xy - I_{y_i}(y^2 + 1) \\ -I_{x_i}(x^2 + 1) - I_{x_i} xy \\ I_{x_i} x + I_{y_i} y \end{pmatrix} \tag{2.7}$$

Then, we can write explicitly [49]:

$$I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + \frac{1}{Z_i} \vec{\mathbf{a}}_i \cdot \mathbf{t} = 0 \tag{2.8}$$

which concisely describes how the image measurements at each pixel $(\vec{\mathbf{a}}_i, \vec{\mathbf{b}}_i, I_{t_i})$ constrain the unknown camera motion $(\mathbf{t}, \boldsymbol{\omega})$ and the scene structure $\frac{1}{Z_i}$. This equation also illustrates the 'scale ambiguity'. There is an inverse relationship between the scale of $\frac{1}{Z}$ and the magnitude of the translation $\mathbf{t}$. Because of this, the instantaneous camera motion $(\mathbf{t}, \boldsymbol{\omega})$ has only five degrees of freedom.

2.4 Direct Solutions for Structure and Motion

The geometric intuition of all direct algorithms computing the motion of the camera is to find parameters $(\mathbf{t}, \boldsymbol{\omega})$ which minimize the deviation from constraint 2.8 defined at every image position. Solving for the structure and motion requires some additional constraint. Otherwise, the set of constraints defined by Equation 2.8 gives a system with n equations and $n + 5$ parameters — a completely independent depth at all n points in an image and 5 parameters of motion.

Negahdaripour and Horn consider one such assumption, of interest in following sections; the case when the scene in view is a plane [49]. For this case, the inverse depth of each pixel i can be expressed as: $\frac{1}{Z_i} = \mathbf{r}_i \cdot \vec{\mathbf{q}}$. Equation 2.8 then becomes

$$I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\vec{\mathbf{q}} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \mathbf{t} = 0 \tag{2.9}$$

This can be extended to solve for scenes where the inverse scene depth is a polynomial function of the image coordinates. To find an estimate of inverse depth that is quadratic, define $\frac{1}{Z_i} = \mathbf{r}_i^T \mathbf{M}_{\mathcal{A}(\mathbf{r}_i)} \mathbf{r}_i$, where $\mathbf{M}$ is an upper-triangular matrix of coefficients to the quadratic surface. Then, each image measurement defines the constraint:

$$\vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\mathbf{r}_i^T \mathbf{M} \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \mathbf{t} = 0 \tag{2.10}$$

This can be generalized to any polynomial model of $\frac{1}{Z}$, but the number of parameters grows quickly and the solution may not be stable for values of $n > 3$ [52].

Furthermore, a single polynomial function is not often an appropriate model of scene depth.

Heel [25] alternately computes motion parameters and structure parameters on successive frames and uses a Kalman filter to predict the depth across multiple frames. This solution has promise if the camera motion does not vary rapidly and if there is no noise in the image measurements. To avoid difficulties with image noise, it is necessary to compute depth at a larger scale than a pixel.

Hanna [22] uses a piecewise planar model of a scene, with fixed patches in the image constrained to lie on the same depth plane; this reduces the number of variables in the depth representation to a function of the number of depth patches rather than the number of image pixels. This model give good results only for scenes that do not have sharp depth discontinuities.

Other solutions make use of the 'positive depth constraint'. It is possible to define $(\mathbf{t}, \boldsymbol{\omega})$ such that the best $\frac{1}{Z}$ at a pixel for Equation 2.8 is negative. This corresponds to a point lying behind the camera. These points are not imaged. Therefore requiring that the computed depth of all visible point is positive gives a constraint on the camera motion. The first methods to enforce this constraint assumed a particular form of camera motion. Horn and Weldon [33] solve for positive depth in the case of only translation, later this was generalized to the case when there is a known bound on the amount of camera rotation [3, 55].

Assuming only that the depth is positive, Fermüller solves the general case for arbitrary camera motions by searching for patterns of normal flow measurements in particular directions. The sign of the normal flow measurements in these patterns is sufficient to find both the translation and rotation [15, 16, 17].

These methods all use constraints on the $\frac{1}{Z}$ values to solve for the structure of the scene and the motion of the camera simultaneously. However, the function they minimize does not have a direct geometric relationship to the scene structure. A sequence of studies has looked more directly at this question, specifically: What happens to the solution for $\frac{1}{Z_i}$ at a pixel when the camera motion estimate is wrong?

What happens is that the estimate $\frac{1}{Z}$ is wrong — and when a scene is reconstructed with the wrong camera parameters, the scene depth is distorted. This distortion has an intricate structure. It is dependent upon the local image patterns, the real camera motion and the scene [4, 5, 11, 19]. Because of this, a scene reconstructed with the incorrect camera motion will be very rugged, with few smooth regions and many discontinuities. Since the world tends to be a set of relatively smooth objects separated by discrete discontinuities, this leads to an intuition for a new constraint. Specifically, search for the motion that minimizes the overall ruggedness of the scene

An implementation of this constraint was formulated by Brodsky [7]. This algorithm divides the image into small patches and searches for the motion that

minimizes the variance of the depth estimate at each patch. It turns out that this constraint can be written as a weighting of the constraints at each pixel. If $\frac{1}{\hat{Z}}$ is the mean inverse depth computed on a patch, then:

$$\frac{1}{Z_i} - \frac{1}{\hat{Z}} = V_i(I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + \frac{1}{\hat{Z}}\vec{\mathbf{a}}_i \cdot \mathbf{t}) \tag{2.11}$$

with the weighting of the constraint function at each pixel:

$$V_i = \frac{1}{\vec{\mathbf{a}}_i \cdot \mathbf{t}} \tag{2.12}$$

There are other weighting functions that give intuitive geometric interpretations. Set $V_i = 1$ everywhere to define the constraint directly related to measurable intensity values. If $V_i = \frac{1}{\sqrt{I_x^2 + I_y^2}}$, then each constraint measures the sum of squared differences of the normal flow and the appropriate projection of the optic flow predicted from the motion parameters. This measure has been used in [44]. These differences are illustrated in Table 2.4.

Enforcing this constraint is efficient and similar, in implementation, to previous work [22, 35, 59] on fitting parameterized models of scene depth or optic flow directly to measurements of image derivatives.

2.5 Ambiguities in Structure from Motion

Chapter 4 is partly an experimental study of ambiguities in the solution for the structure and motion of the scene. Previous work on the ambiguity inherent in the problem has typically considered the case where there are correspondences

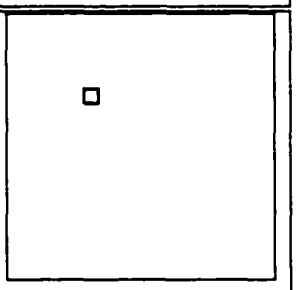
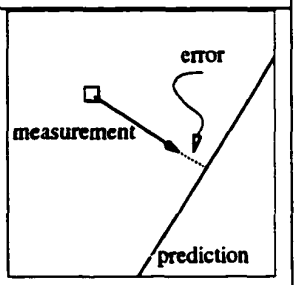
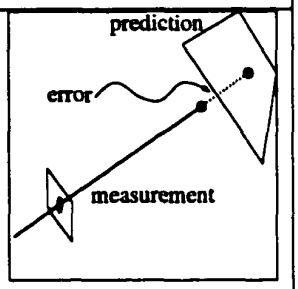
V_i	Measurement	Prediction	Pictorially
1	measured valued of I_t	prediction of I_t based on camera motion, measurement of I_x, I_y and depth model	
$\frac{1}{\sqrt{I_x^2+I_y^2}}$	normal flow magnitude computed from measurement of I_t, I_x, I_y	normal flow magnitude predicted by camera motion and depth model	
$\frac{1}{\vec{a}_i \cdot \vec{t}}$	depth computed from measurement of I_t, I_x, I_y and camera motion	depth predicted by depth model.	

Table 2.1: Summary of geometric meaning for natural error functions

between the images. This work can be divided into absolute ambiguities, cases where there is no possible way to distinguish between multiple interpretations of the scene, and noise based ambiguities, situations which become ambiguous given that the image points are measured with some error.

There exists a set of surfaces for which even a zero noise motion field may be ambiguous. That is, there exists camera motions and surfaces such that, a different camera motion, and a different surface can lead to exactly the same image measurements. These two interpretations cannot be distinguished. These surfaces are hyperboloids of one sheet [30], and the ambiguity only exists if the camera lies also on the hyperboloid.

The general case of ambiguity caused by errors in optic flow measurements has received considerable attentions for a variety of assumptions; orthographic projection and perspective projection, two frames or many frames, and, in the case of many frames, smoothly varying motions or arbitrary motions (for example [69, 2, 50, 67]). In general, the ambiguity is correlated with the following factors: small field of view, distance to the scene, small camera translation, sparse flow measurements and correlated noise between flow measurements [2]. The ambiguity is sometimes called the "bas-relief" ambiguity because the reconstructed 3D points are scaled linearly in the depth dimension.

There have also been studies of ambiguities in structure from motion without using explicit point correspondences. Brodsky has studied the absolute ambigu-

ities when the input data is only the direction (not the magnitude) of the flow field. If more than half of the viewing sphere is imaged, there is no ambiguity. Otherwise, the only ambiguity arises if the error in the estimated translation is perpendicular to the error in estimated rotation and the depth of the scene lies between a particular second and third order surface [8]. Fermüller has studied the negative depth constraint using normal flow as the basic image measurement. In the case of noise in these measurements, the motion estimates that minimize the negative depth estimates are constrained so that the translational estimate lies on a line connecting the image center to the real translation, and the rotational error is perpendicular to the translational error. When the entire sphere is imaged, given a rotational error, the optimal translation is the correct one; given a translational error, the optimal rotational error is again perpendicular [18].

Chapter 3

Estimation of Optic Flow

3.1 Introduction

The set of all the image intensity measurements $I(x, y, t)$ is exactly the information recorded in a video sequence. However, this set of pixel intensities is not closely related to either the camera motion or the scene structure. Other representations of how the images are changing are easier to analyze. The most common approach computes two-dimensional image measurements which correspond to the velocity measurements of image patterns, called optical flow. This chapter demonstrates that optic flow is very hard to estimate; common approaches to calculate the optic flow vector field give a bias. This bias is related to the local image structure, the true motion, and the distribution of error in the local image measurements. To estimate, or correct for the bias, it is necessary to estimate this error distribution, a task that is especially difficult for sensors that move

through natural scenes. The form of this bias also explains a number of perceptual illusions. The algorithms that are presented in subsequent chapters use only the derivatives of the image intensity function, a representation of image change that does not suffer from this bias.

3.2 Computing Optic Flow

The optical flow field represents an approximation of the projection of the field of motion vectors of the 3D scene points on the image. Computational considerations as well as biological measurements suggest that optical flow is derived in a two-stage process[1, 66]. In a first stage, from local image measurements, the velocity component perpendicular to linear features is computed. The situation is illustrated in Figure 3.2. The velocity vector of a one-dimensional feature (such as a line or piece of contour) viewed through a small aperture is inherently ambiguous, as it is consistent with any vector falling on the constraint line [63]. Only the velocity component perpendicular to the feature in the direction of the motion is well defined. In the computational literature this component is referred to as normal flow and the ambiguity is referred to as the aperture problem[42].

In a second stage, normal flow measurements from features in different directions residing in a neighborhood are combined in order to derive the complete optical flow. The combination of flow vectors, however, constitutes an intricate computational problem. Computational problems arise at the locations of flow

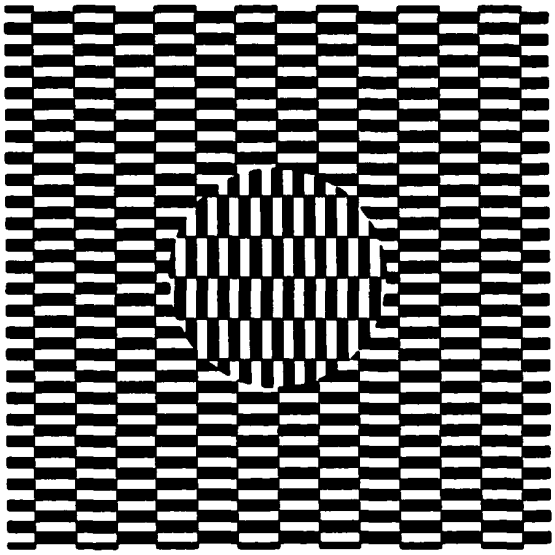


Figure 3.1: A pattern similar to one by Ouchi [51]. Slight motion of the page will give a distinct impression of independent motion.

discontinuities [29, 32], which are due to objects at different depths or differently moving scene elements. Within small image patches arising from coherently moving, smooth parts of the scene, the optical flow field is well approximated with a parametric model varying, for example, as a constant, linear or quadratic function of the image coordinates [6]. At the locations of discontinuities, however, it is not, and if image measurements across discontinuities are combined, very erroneous optical flow measurements may be derived [32]. To avoid the smoothing over boundaries, knowledge of where the discontinuities are seems to be necessary, which is difficult to obtain from local image measurements.

Even within areas of smooth flow, the computation of optical flow poses a problem. The focus of this chapter is to show that for statistical reasons it is very difficult to obtain accurate optical flow estimates. The ideas underlying

the statistical explanation of optical flow estimation are as follows: Local one-dimensional flow components — normal flow measurements — are estimated with error. We assume that the estimates of these components are unbiased. However, when combining the one-dimensional measurements in a neighborhood an estimate of optical flow is obtained which is biased. The estimated value depends on the distribution of image gradients, the actual flow, and the error in the normal flow.

The statistical model explains a number of psychophysical findings, which are concerned with the perception of motion in patterns with a sparse, limited set of spatial frequencies. The gradient distribution in the patterns is such that the bias is highly pronounced. In particular, we elaborate on the Ouchi illusion and related experiments [27, 28, 38, 37]. The Ouchi illusion, as shown in Figure 3.1, consists of two black-and-white rectangular checkerboard patterns oriented in orthogonal directions – a background orientation surrounding an inner disc. Scanning eye movements over these patterns generate the striking perception of relative movement of the inner disc. Our explanation lies in the estimation of differently biased flow vectors in the two patterns which in turn give rise to different 3D motion estimates that cause one pattern to move relative to the other. Furthermore, we explain a number of observations found in the study of moving plaids. These two-dimensional patterns consist of two one-dimensional gratings (with sine, cosine, or rectangular underlying waveforms) with different

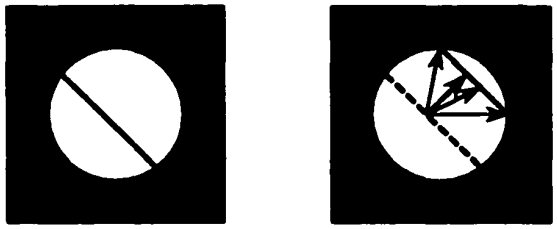


Figure 3.2: Aperture problem: (a) Line feature observed through a small aperture at time t. (b) At time $t + \delta t$ the feature has moved to a new position. It is not possible to determine exactly where each point has moved to. From local measurements only the flow component perpendicular to the line feature can be computed.

orientations whose motion appears coherent or incoherent depending on various parameters such as contrast, speed, and spatial frequency. They were introduced originally by Adelson and Movshon in [1] and have since then been studied extensively to assess models used in the explanation of human flow computation [34, 68]. The combination of measurements of patterns different than plaids has sought to determine when smooth contours are seen to move rigidly as opposed to non-rigidly [47, 48].

Models used in the computational and biological literature to estimate the optical flow in a two-stage process can be placed in roughly two categories, those modeling computations in image space and those in spatiotemporal frequency space. The modeling conducted in this chapter concentrates on the first category.

In image space models, the one-dimensional motion component of features is estimated by assuming the conservation of image intensity or some function of it. The gradient based approaches assume that image intensity does not change

over a small time interval. Denoting the image intensity as I, its spatial (in x and y direction) and temporal derivatives as I_x, I_y, and I_t, respectively, and the velocity of image points in x and y direction as u and v, the following constraint is obtained:

$$I_x u + I_y v + I_t = 0 \tag{3.1}$$

This equation, called the optical flow constraint equation [31], defines the component of flow in the direction of spatial gradient (I_x, I_y)—the normal flow. Other, more elaborate techniques consider functions of the image intensity or the local intensity distribution to be conserved. In order to derive the optic flow from the normal flow measurements in a neighborhood, a second constraint has to be invoked. Usually it is assumed that optical flow varies smoothly. This is achieved by either modeling the flow field explicitly as a polynomial in the image coordinates, or modeling the smoothness through some function in the derivatives of the flow values leading to a regularization formulation [26, 29, 53].

The estimation of flow then amounts to an optimization problem minimizing some function of deviation from the model; usually, a least squares minimization is used. The intersection of constraints (IOC) model often used in the psychological literature is a typical instance of a smoothness constraint. It assumes the optical flow to be constant within a neighborhood, an assumption that is justified within small regions, or if the motion in view originates in a translation due to a fronto-parallel plane.

In the modeling conducted here, we employ the optical flow constraint equation and assume constant flow within a neighborhood. As the psychological experiments analyzed in this chapter are concerned with translations in the frontoparallel plane, this model is appropriate and simplifies the exposition. For combining normal flow vectors into optical flow, we use the least squares estimation model. We will show, however, that the bias found in the estimation of flow is not due to the particular models employed, rather it is inherent in the geometry of the constraints placed on combining one-dimensional motion components into optical flow.

The remainder of this chapter is organized as follows: In Section 3.3, we discuss the psychophysical studies detailing the perception of the Ouchi illusion and related biases in the perceptions of plaid motion. In Section 3.4, we analyze an IOC type model to compute estimates of patch velocity directly from noisy measurements of image derivatives. We then discuss the bias and provide graphic illustrations of the estimated flow for the patterns of limited sets of gradient directions occurring in the psychophysical stimuli. In Section 3.5, this analysis is used to explain why local patch velocities do not combine to form a coherent perception of pattern motion in the Ouchi illusions and related patterns, and also to explain both coherence judgments and directional biases in the perception of plaid motion. Section 3.6 is devoted to a general discussion of statistical techniques proposed in the literature on estimation theory to deal with the noise

model used in the analysis and the inherent problems in applying these techniques to the problem of optical flow estimation. As will be shown, correcting the bias would require knowledge of noise not attainable from a limited set of measurements of the particular psychophysical stimuli—thus demonstrating that the bias is not a peculiarity of the particular computational models we employ.

3.3 Psychophysical Experiments on Motion Perception

The striking illusion discovered in 1977 by the graphic artist H. Ouchi is evoked by a stationary picture which consists of a checkerboard pattern superimposed on another rectangular checkerboard oriented in orthogonal direction (Figure 3.1). Small retinal motions, or slight movements of the paper, evince a segmentation of the inset of the pattern, and motion of this inset relative to the surround. The illusion remains under a variety of viewing distances and angles. Some observers report an apparent depth discontinuity, with the center floating as it moves atop the background [58]. Here, we summarize the findings from psychophysical experiments which have studied this illusion specifically, and then continue with results of plaid experiments which attempt to find general parameters of how local flow measurements are combined.

Khang and Essock [38, 37] performed experiments with a number of variations of the original pattern to evaluate the impact of various parameters, such as orientation and size of the pattern elements, luminance, and blurring, on the perceived strength of the illusion. In most of the figures they used a simplified version of the illusion with just a one-dimensional square wave grating present in the inset. We concentrate here on the first set of experiments in [38] conducted with only two-dimensional patterns. In these experiments they replaced the periodic rectangular checkerboard patterns in the inset and surround by various other 2D periodic patterns, each composed of two 1D functions, one of a short frequency and one of an orthogonal longer frequency.

The particular patterns used, namely the original rectangular checkerboard, a sinusoidal, a trapezoidal, a triangular, a sawtooth and an added sinusoidal pattern are described and shown in Figure 3.3. Subjects were asked to view the patterns freely and rate the magnitude of the apparent motion; the results of their findings are displayed in Figure 3.4.

The second set of studies [27, 28] used a simplified stimulus replacing the 2D patterns in the inset and surround with sinusoidally modulated contrast gratings of the same spatial frequency. The two gratings as shown in Figure 3.5a were tilted symmetrically about the vertical axis: $\theta/2$ degrees to the left and right, respectively. To give the illusion, this stimulus was presented moving vertically on a computer screen, which can be simulated through vertical up-and-down

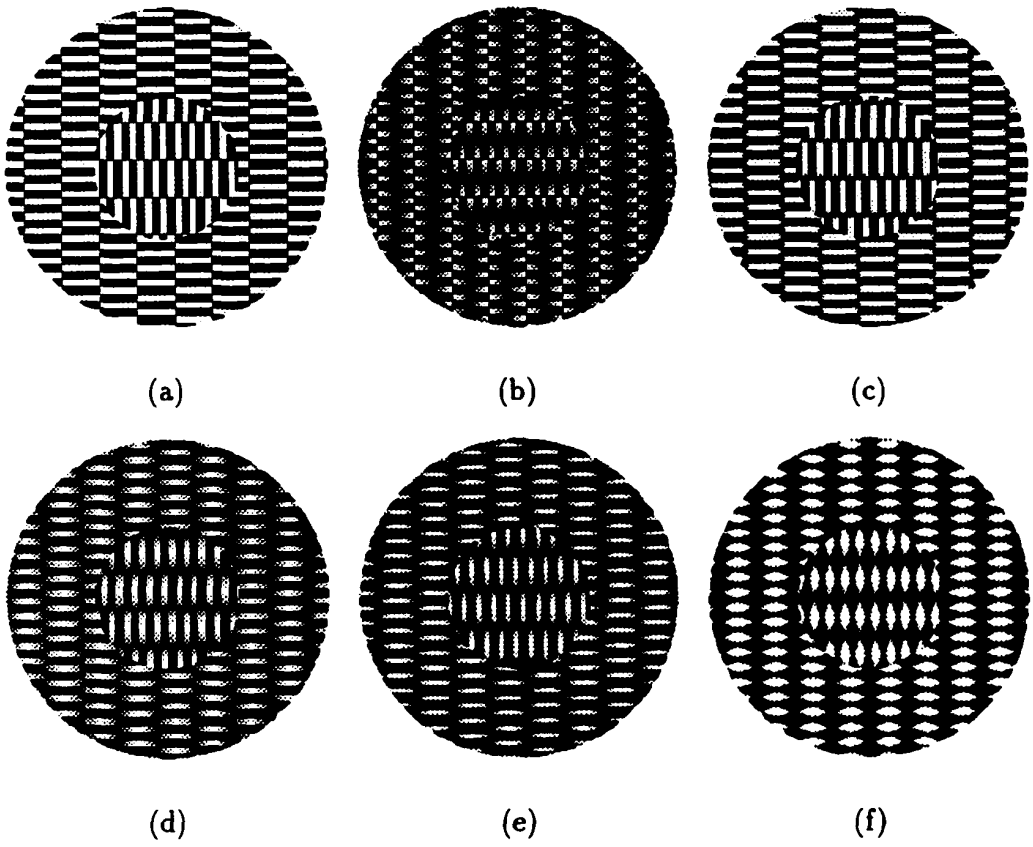


Figure 3.3: Variations of the Ouchi pattern used in [38]. Patterns were formed by combining two one-dimensional periodic functions. (a) Rectangular checkerboard composed by multiplying a horizontal square-wave and a vertical square-wave function. (b) Sawtooth pattern composed of the product of a sawtooth-wave and a square-wave function. (c) Trapezoidal pattern composed of the product of a trapezoidal-wave and a square-wave function. (d) Triangular pattern composed of the product of a triangular wave and a square-wave function. (e) Sinusoidal pattern composed of the product of a horizontal sine wave and a vertical sine wave function. (f) Added sinusoidal pattern composed by adding a horizontal sine-wave and a vertical sine-wave function.

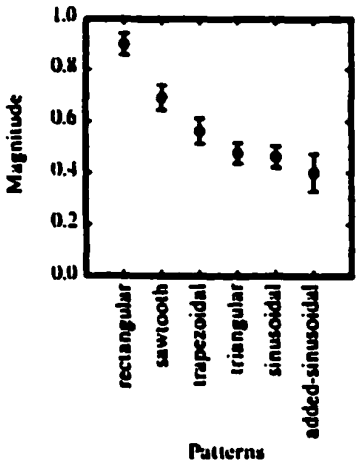
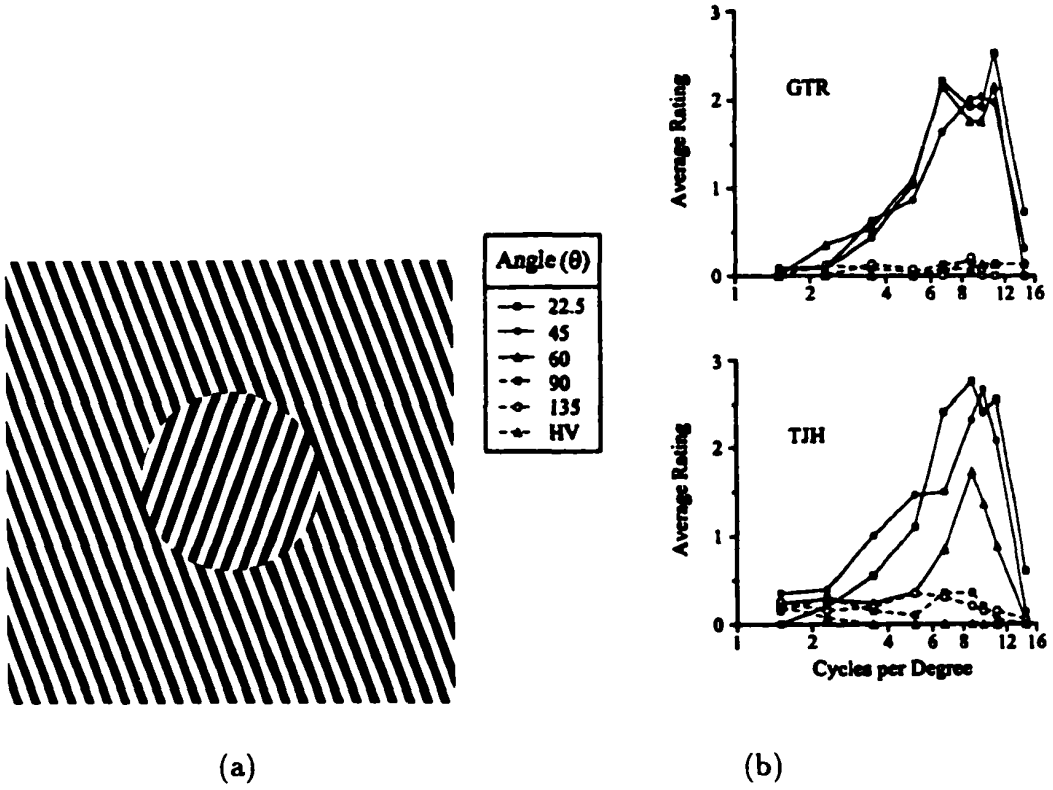


Figure 3.4: Means and standard errors of the magnitude of the motion illusion as a function of the six different 2D patterns [38].

movements of the paper. The apparent motion of the inset is seen orthogonal to the grating orientation, oriented in the direction whose angle with the overall motion of the paper is less than 90°.

The parameters, which they varied in their figures, were the spatial frequencies and the angle between the gratings. With short presentation times preventing the possibility of tracking motions, they found that the strength of the illusion of relative motion decreases with the angle between the gratings, and strong responses only for angles smaller than 90 degrees and frequencies between 6–12 cycles/degree as shown in Figure 3.5b.

As a possible explanation for the illusion, Hine et al. [27, 28] suggest an anomaly of the visual system in integrating local velocity signals into a rigid percept – component motion vectors that differ in direction by more than 120° stimulate entirely different grating cells and motion channels and are not combined.



(a) (b)

Figure 3.5: (a) Reduced stimulus in experiments used in [27, 28]. The surround and inset gratings were tilted symmetrically about the vertical meridian, each by an angle $\theta/2$ from this meridian. (b) Strength of the relative illusion (evaluated by average ratings) as a function of the angle θ and spatial frequency of the stimulus, plotted for each observer. In HV the inset was vertical and surround horizontal. The acute angles ($\theta < 90°$) produced a greater illusory effect than obtuse angles.

Khang and Essock [37, 38], cite as a possible cause the interactions between spatially overlapping ON and OFF units — in particular, saturation and non-linear response profiles of visual channels responsive to brightness changes leads to an overall impression of motion. In particular, there is (a) a non-linear response of channels responding to luminance change over time, and (b) the visual system cannot report accurate local pattern intensity if contrast reversals occur abruptly over large parts of the image. These are distortions of the spatial and temporal image intensity derivatives — a formal analysis of the effects of these errors on the integration of local flow measurements is the focus of this chapter.

The integration of local velocity signals has been studied extensively in the context of understanding the perception of moving plaids. Plaids are combinations of two wave gratings of different orientations each moving with a (typically identical) constant speed. For any such "moving plaid", there is always some planar velocity the whole pattern can undergo which would produce exactly the same retinal stimuli. However, for particular variations of the spatial frequencies of the component gratings, their relative orientations, contrasts or speeds, human perception is of two separate motions, with one grating "sliding" over the other. In particular cases, one can perceive a constant, coherent motion of the pattern, biased away from the unique velocity which would account for all retinal signals.

Whether or not a particular plaid pattern is judged to be coherent depends upon the contrast, spatial frequency and motion directions of the components [1].

If the component motion directions differ by more than 90 degrees, the plaid motion is not perceived as coherent for a variety of contrast and spatial frequency conditions [39]. A plaid made of high contrast orthogonal gratings is perceived as coherent despite spatial frequency differences of up to 3 octaves [56]. If the component gratings are square waves, the luminance profile of the intersection regions also changes the perception of coherence, independently of other changes in spatial frequency or contrast [60].

The motion of a coherent plaid pattern can be theoretically computed using the intersection of constraints model (IOC) — the vector component obtained from each individual grating constrains the local velocity vector to lie upon a line in velocity space, the intersection of the lines defines the motion of the plaid [1]. Some plaid patterns are perceived as coherently moving, but with a velocity different than that predicted by the IOC model. This bias affects both the direction and magnitude of the perceived velocity. The velocity of plaid patterns made of different one dimensional gratings is biased towards the grating of higher contrast [60, 40]. For type 2 plaids, where the IOC velocity is not between the component directions [20], the bias is towards the average of the component vectors [21, 10]. Plaids comprised of gratings of different spatial frequencies are also biased in both direction and length [57, 40]. In no case is the there an overestimate of the plaid velocity compared to the IOC prediction.

Monte-Carlo experiments have attempted to determine the expected value and variance of velocity calculated with the IOC method, for the case where the one dimensional motion is measured with some Gaussian distributed error [21, 47]. Both experiments proceeded by generating a speed measurement for each component direction, corrupting this measurement with Gaussian noise, and then computing the IOC prediction from this pair of constraints. The distribution of estimates created in this method is not biased away from the IOC motion [21], and the variance of these estimates is correlated with accuracy of directional perception [47]. The next section extends the analysis of the IOC model to accept more than two noisy local motion measurements, and finds a bias that is dependent on the distributions of the local orientations.

3.4 Analysis of Optical Flow Estimation

We analyze the estimation of optical flow from local measurements of changes in the image intensity using least squares minimization of the optic flow constraint equation. We assume that the flow is constant within the region of gradient measurements. As input we consider a set of estimated spatial and temporal gradient measurements $(\hat{I_{x_i}}, \hat{I_{y_i}}, \hat{I_{t_i}})$ which are compounded of the actual values

$(I_{x_i}, I_{y_i}, I_{t_i})$ and noise $(n_{x_i}, n_{y_i}, n_{t_i})$.

$$\hat{I_{x_i}} = I_{x_i} + n_{x_i} \tag{3.2}$$

$$\hat{I_{y_i}} = I_{y_i} + n_{y_i} \tag{3.3}$$

$$\hat{I_{t_i}} = I_{t_i} + n_{t_i} \tag{3.4}$$

with

$$\hat{\mathbf{I}}_s = \begin{bmatrix} \vdots & \vdots \\ \hat{I_{x_i}} & \hat{I_{y_i}} \\ \vdots & \vdots \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{I}}_t = \begin{bmatrix} \vdots \\ \hat{I_{t_i}} \\ \vdots \end{bmatrix} \tag{3.5}$$

The optical flow constraint equation relates the locally image intensity derivatives to the image velocity. Assuming that the optical flow $\mathbf{u} = (u, v)$ is constant within the region considered, it thus is described by the following over-determined system of equations:

$$\hat{\mathbf{I}}_s \mathbf{u} + \hat{\mathbf{I}}_t = 0 \tag{3.6}$$

Solving (3.6) by a standard least squares estimation for the flow $\mathbf{u}$ yields

$$\mathbf{u} = -\left(\hat{\mathbf{I}}_s^T \hat{\mathbf{I}}_s\right)^{-1} \hat{\mathbf{I}}_s^T \hat{\mathbf{I}}_t \tag{3.7}$$

We consider the effects of the following noise model. The measurement of each image derivative is corrupted by an additive error, these errors are zero mean Gaussian random variables, independent at different image locations, but with

possible dependencies between the spatial and temporal derivatives at one location. The second moments of such noise are simply described through a covariance matrix, with one remark. As the model should provide measurements which are symmetric with respect to reflections along the coordinate axes, we assume the noise component due to correlation between the spatial and temporal derivatives dependent in sign on the sign of the derivatives. If one of the derivatives is positive and the other is negative, such as in the first quadrant, we assume positive correlation, otherwise we require sign change. This kind of noise would result if the derivative operations are carried out by a symmetric set of unidirectional derivative operators which are activated selectively depending on the sign of the gradients, and thus collectively performing either forward or backward differentiation.

To obtain a more compressed notation, we also assume the noise in the two spatial components is equal. This might be an oversimplification for real systems, but it does not affect the forthcoming analysis. This means that the variances and covariances of the noise components are given as:

$$E(n_{x_i}) = E(n_{y_i}) = E(n_{t_i}) = 0$$

$$E(n_{t_i}^2) = \sigma_t^2, E(n_{x_i}^2) = E(n_{y_i}^2) = \sigma_s^2$$

$$E(n_{x_i} n_{y_i}) = 0$$

$$E(n_{x_i} n_{t_i}) = \sigma_{xt} = -\text{sgn}(I_{x_i} I_{t_i}) \cdot \sigma_{st}$$

$$E(n_{y_i} n_{t_i}) = \sigma_{yt} = -\text{sgn}(I_{y_i} I_{t_i}) \cdot \sigma_{st}$$

In the absence of error in the spatial gradient measurements $\hat{\mathbf{I}}_s$, standard least squares methods give an unbiased estimator. The expected value $E(\mathbf{u})$, obtained from (3.7), corresponds to the true optical flow $\mathbf{u}_0$.

However, errors in this measurement matrix can lead to a bias such that the expected value of the estimated flow $\hat{\mathbf{u}} = E(\mathbf{u})$ is no longer the true optical flow. The form of this bias is apparent in the second order Taylor expansion of the expected value of the least squares solution as a function of the variance and covariances of the noise in the measurement matrices. The first order terms vanish, the only non zero terms that remain in the expansion at zero noise ($\mathbf{n} = 0$) are:

$$\hat{\mathbf{u}} = \mathbf{u}_0 + \sum_i \frac{\partial^2}{\partial n_{(x,y,t)_i}{}^2}(\hat{\mathbf{M}}^{-1}\hat{\mathbf{b}})\rfloor_{\mathbf{n}=0} \frac{\sigma^2_{n(x,y,t)_i}}{2} + \sum_i \frac{\partial^2}{\partial n_{x_i} dn_{t_i}}(\hat{\mathbf{M}}^{-1}\hat{\mathbf{b}})\rfloor_{\mathbf{n}=0}\sigma_{n_{x_i} n_{t_i}} + \sum_i \frac{\partial^2}{\partial n_{y_i} dn_{t_i}}(\hat{\mathbf{M}}^{-1}\hat{\mathbf{b}})\rfloor_{\mathbf{n}=0}\sigma_{n_{y_i} n_{t_i}} \quad (3.8)$$

where $\hat{\mathbf{M}} = \hat{\mathbf{I}}_s^T\hat{\mathbf{I}}_s$, and $\hat{\mathbf{b}} = \hat{\mathbf{I}}_s^T\hat{\mathbf{I}}_t$.

Algebraic manipulation of the above derivative leads to an expression of $\hat{\mathbf{u}}$ which can be written as a sum of three components: the true optical flow $\mathbf{u}_0$, a component which is due to the variance in the spatial derivative noise only (which we refer to as variant noise), and a component which originates from the covariance terms of the noise in the temporal and spatial measurements (which we refer to as covariant noise). The exact expression is given in the appendix; its dominant factors are

$$\hat{\mathbf{u}} = \mathbf{u}_0 - K_1 \left(\sum_i \mathbf{M}^{-1}\mathbf{u}_0 \right) - \sum_i K_{2_i} \mathbf{M}^{-1} \begin{bmatrix} \mathrm{sgn}(\sigma_{xt_i}) \\ \mathrm{sgn}(\sigma_{yt_i}) \end{bmatrix} \tag{3.9}$$

with $K_1 = \sigma_s^2$ and $K_{2_i} = \frac{\sigma_{st}}{\sigma_s\sigma_t} \cdot (\sigma_s^2 + \sigma_t^2 + \sigma_s\sigma_t + 2\frac{\sigma_{st}^2}{\sigma_s\sigma_t} + (\frac{\sigma_{xt_i}}{\sigma_s\sigma_t}u + \frac{\sigma_{yt_i}}{\sigma_s\sigma_t}v)(\sigma_t^2 + 2\sigma_s\sigma_t))$ and $\mathbf{M} = \mathbf{I}_s^T\mathbf{I}_s$, the matrix of uncorrupted spatial gradient values.

The effect of the gradient distribution on the bias of the computed flow can be interpreted through its effect on the matrix $\mathbf{M}^{-1}$. In the case of a uniform distribution of image gradients in the region where flow is computed, $\mathbf{M}$, (and therefore $\mathbf{M}^{-1}$) are multiples of the identity matrix, leading to a bias solely in the length of the computed optical flow. Both the variant term and the covariant term lead to an underestimation in the length. In a region where there is a unique gradient vector, $\mathbf{M}$ will be of rank 1, this is the aperture problem. In the general case, the bias can be understood by analyzing the eigenvectors of $\mathbf{M}$. As $\mathbf{M}$ is a real, symmetric matrix, its two eigenvectors are orthogonal to each other with the direction of the eigenvector corresponding to the larger eigenvalue dominated by the major direction of gradient measurements. $\mathbf{M}^{-1}$ has the same

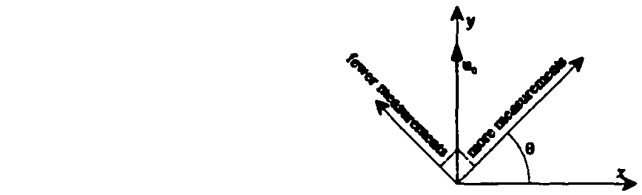
eigenvectors as $\mathbf{M}$ and inverse eigenvalues. Thus, the eigenvector corresponding to the larger eigenvalue of $\mathbf{M}^{-1}$ has a direction dominated by the normal to the major orientation of image gradients, and the product of $\mathbf{M}^{-1}$ with any vector is most strongly influenced by this orientation. This effects the variant term to lead to an underestimation in the length of the optical flow and a bias in direction toward the major direction of gradients. The covariant term in most cases also leads to an underestimation in the length and its influence on the direction can be either way, toward or away from the major direction of gradients, depending on the particular gradient distribution.

The following figures illustrate the bias. Figure 3.6 displays the expected values of the noise terms for a gradient distribution as it occurs in one of the regions of the Ouchi illusion shown in Figure 3.1 with blocks four times longer than they are wide. In particular, image gradients are in two orthogonal directions with four times as many measurements in the one direction as in the other. The actual optical flow is along the positive y axis and of length one and the plots show the change in the bias as the relative angle between the perpendicular gradients and the true flow direction varies. The angle θ is measured between the positive x axis and the direction of more gradients; the other gradient direction forms an angle $\theta + \pi/2$ with the x axis (see Figure 3.6a). Figures 3.6(b,c) show the error in length and angle due to the variant term and Figures 3.6(d,e) show the same

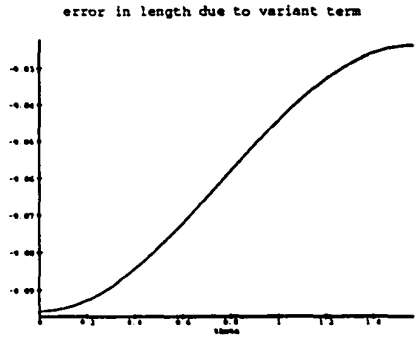
errors for the covariant terms. The plots are based on the exact second order Taylor expansion as given in the appendix.

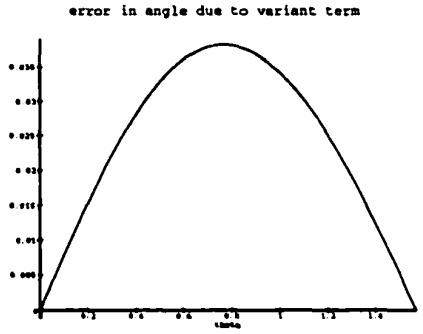
For such gradient distribution the bias can be understood rather easily. The eigenvectors of $\mathbf{M}$ are in the directions of the two gradient measurements with the larger eigenvalue corresponding to the more gradients. As $\mathbf{u_0} = (0, 1)$, the variant term in (3.9) leads to a bias in length as shown by the curve in Figure 3.6b, which takes its minimum at 0 and maximum at $\pi/2$ (that is, when $\mathbf{u_0}$ is aligned with the major gradient direction). The error in angle is greatest for $\pi/4$ (that is, when $\mathbf{u_0}$ is exactly between the two eigenvectors of $\mathbf{M}^{-1}$ and it is 0 for 0 and $\pi/2$ (Figure 3.6c). Overall, this means the bias due to the variant term is largest when the major gradient direction is normal to the flow and is nearly eliminated when it is aligned with the flow, that is, in the Ouchi pattern, when the long edge of the block is perpendicular to the motion. It is always negative in length and towards the major gradient direction.

Regarding the covariant term, as K_{2_i} in (3.9) is constant within the range of 0 to $\pi/2$ and the vectors $(\text{sgn}(\sigma_{xt_i}), \text{sgn}(\sigma_{yt_i}))$ are along the first and second meridian, the covariant term can be written as $K\mathbf{M}^{-1}\mathbf{a}$, with K a positive constant and $\mathbf{a} = 4(1, 1) + (-1, 1) = (3, 5)$. This leads to error functions, as shown in Figures 3.6(d,e) which appear to be shifted to the left of the θ axis with regard to the variant bias. This bias is always negative in length and mostly toward the minor gradient direction. The bias for angles θ between $\pi/2$ and π is obtained

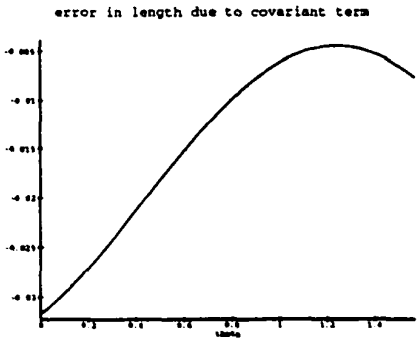


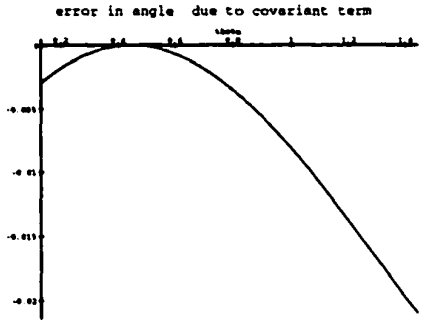
(a)





(b) (c)





(d) (e)

Figure 3.6: (a) 16 measurements are in a direction at angle θ from the x axis and 4 measurements are in the direction $\theta + \pi/2$. The optical flow is along the positive y axis and of length one. (b) Expected error in length of variant term. (c) Expected error in angle due to variant term measured in radians between the expected flow and the actual flow. (d, e) Expected error in length and angle for covariant term. The error has values $\sigma_s = \sigma_t = 0.15$ and $\sigma_{st} = 0.1 \cdot \sigma_s^2$.

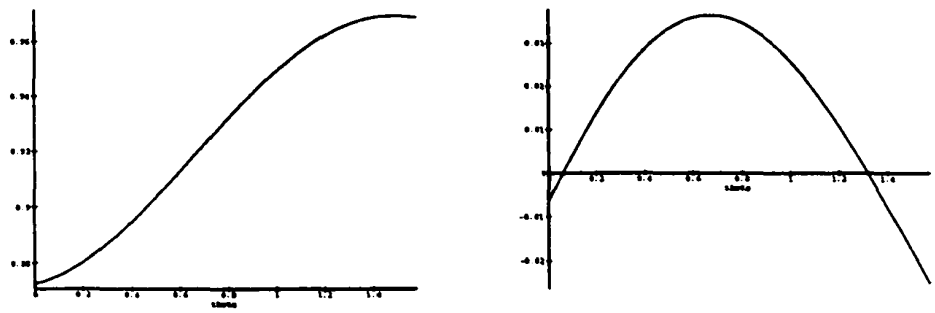


Figure 3.7: Expected length of optical flow and expected error in angle for the gradient distribution and error terms of Figure 3.6.

from the above plots by reflecting the curves upon $\pi/2$ and changing the sign for the error in angle (such that the variant bias is always toward the major and the covariant term mostly toward the minor direction).

To see the combined effect of the error terms, Figures 3.7(a,b) show the expected length of the estimated flow and the error in angle for the same configuration as in Figure 3.6. As we expect the covariant to be much smaller than the variant term, the graph is mostly determined by the latter.

Figure 3.8 illustrates the measurements of component motions in the reduced Ouchi stimulus or symmetric plaids. The two gradient directions are symmetric with regard to the y axis (the direction of motion) with the angle φ measured between the direction of more gradients and the x axis (Figure 3.8a). The correct pattern motion is of length one, and the bias is shown for a receptive field which has more measurements in rightward component direction, simulating a receptive

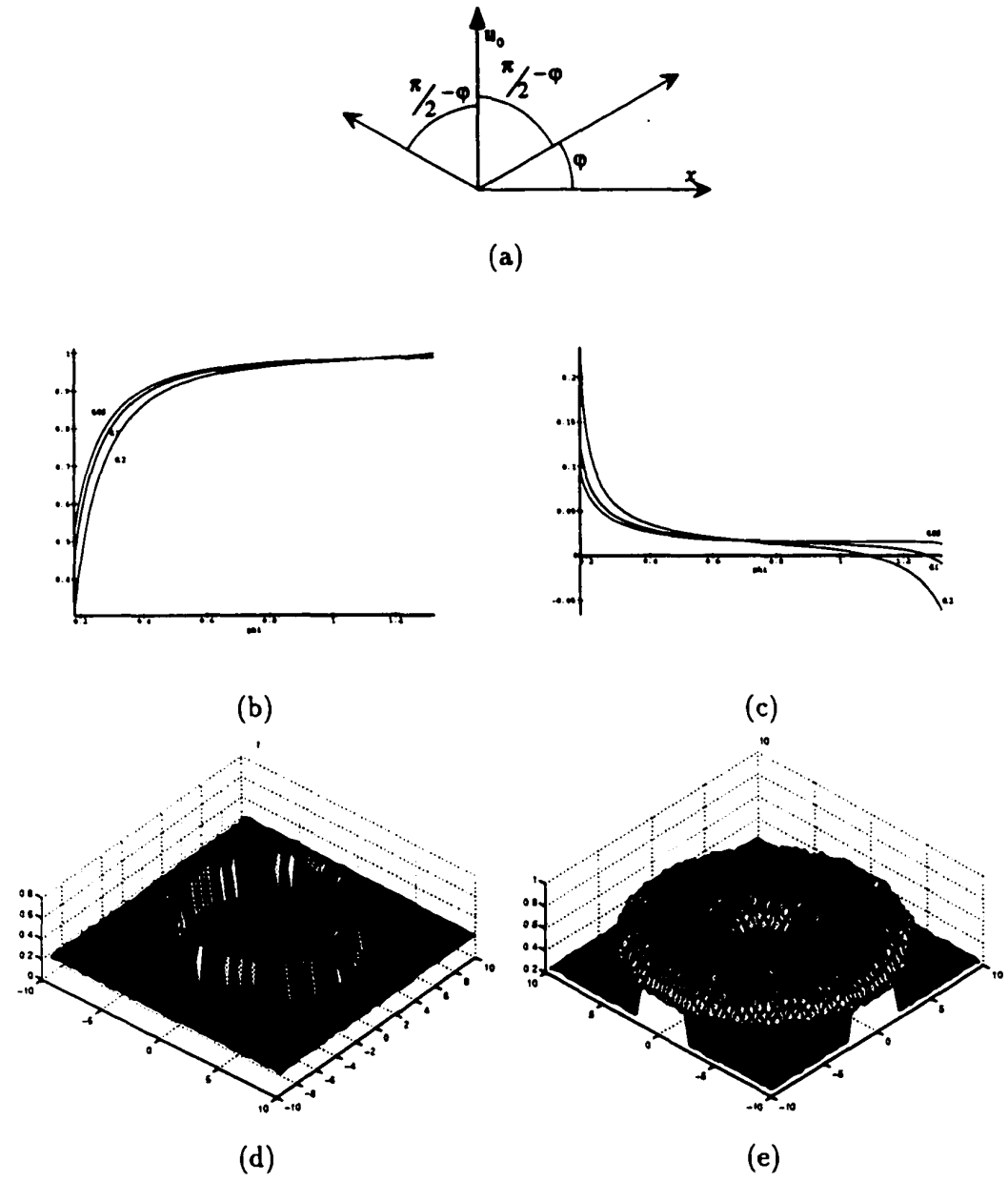
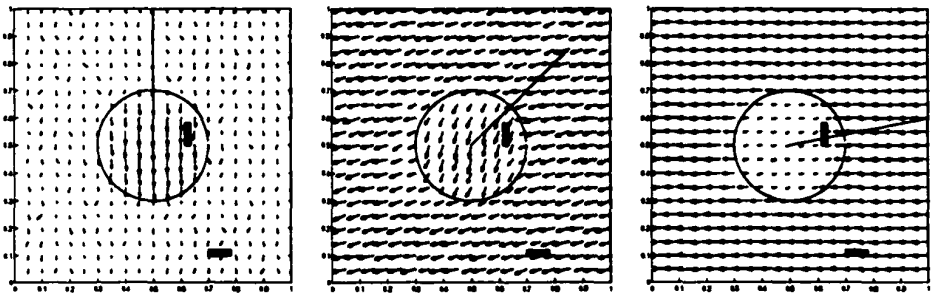


Figure 3.8: (a) Gradients motions in symmetric diagonal directions; For a receptive field with four times as many measurements in the rightward component direction, (b, c) give the error in length and angle due to both variant and covariant term. The noise has values $\sigma_s = \sigma_t = 0.1$ and $\sigma_{st} = 0.1 \cdot \sigma_s^2$. (d, e) show the residual error of the least squares solution, when combining measurements in a small receptive field (d) and a much larger receptive field (e).

field position to receive more input from the inset of the pattern than the outset. The same relative number of measurements would come from a plaid pattern made from components of different frequency. Figure 3.8 (d,e) show the residual of the least squares solution for different receptive field sizes – this residual is much smaller in regions where the receptive field only gets input from one component direction. A small receptive field gives a more clear demarcation of where no optic flow fits well with the measured constraints, it is more difficult to localize the boundary with a broader receptive field.

3.5 Explanation

The previous analysis underlies the nature of the Ouchi illusion. The relative angles between the real motion and the predominant gradient direction differ in the inset and the surround, so the regional velocity estimates are biased in different ways. When, instead of freely viewing the pattern of Figure 3.1, the page is moved in different directions, we observe that the illusory motion of the inset is mostly a sliding motion orthogonal to the longer edges of the rectangle and in the same general direction as the motion of the paper. Using Figure 3.7, it can be verified that the projection of the vector resulting from the difference of the bias vector in the inset and the bias vector in the surrounding area is, for almost all angles in this direction. For example, when the motion is along the first meridian (to the right and up), the bias in the inset is found in the graph at angle



(a) (b) (c)

Figure 3.9: The residual regional motion vector field. The vectors shown are the difference between the true motion and the calculated motion. One block is shown to show the relative orientation for the inset and the outset of the illusion, the width of the block gives the relative number of vertical and horizontal gradient measures. The line from the center is the direction of the true motion. The noise is Gaussian and the spatial gradient magnitude is one. In (a) and (b), $\sigma_s = \sigma_t = 0.1$ and there is no covariance; in (c) $\sigma_s = \sigma_t = 0.2$ and $\sigma_{st} = 0.2 \cdot \sigma_s^2$.

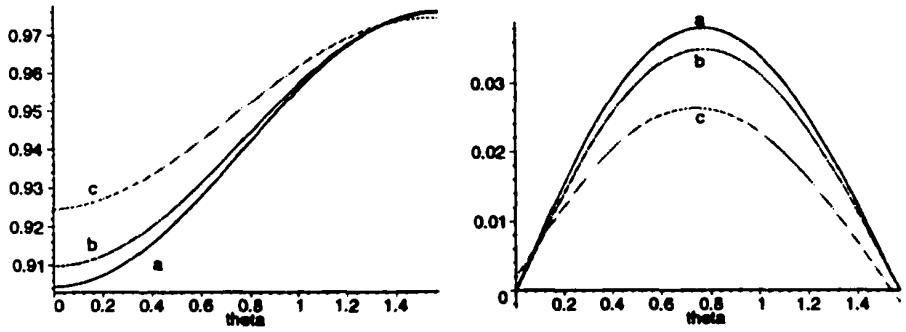
$\theta = \pi/4$ and in the outset at $\theta = 3\pi/4$. The two bias vectors are of about the same length, each in direction towards the gradients of the longer edge, and the resulting projection of difference vectors is to the right. If the motion of the paper is to the right, the difference in bias vectors is mostly due to length resulting in a perceived motion to the right, and if the motion of the paper is upwards, the difference vector is downwards. Its projection on the major gradient direction of the inset is close to zero and thus hardly any illusory motion is perceived. Figure 3.9 shows, for a set of true motions, the biases in the perceived motion. The three bias fields were created with a variety of noise magnitudes, receptive field sizes, and covariance between the noise in the gradient measurements to show the robustness of the effect.

We assume that in addition to computing flow the system also performs segmentation, which is why a clear relative motion of the inset is seen. When experiencing the Ouchi illusion under free viewing conditions, the triggering motion is due to eye movements which can be approximated through random, frontoparallel translations. As the difference in the bias vectors of the inset and surround has a significant projection on the dominant gradient direction of the inset for a large range of angles (that is, directions of eye movements) the illusion is easily experienced.

In the figures of Khang and Essock [38], patterns were used which have more than just two spatial gradient directions. From the rectangular to the sawtooth, the trapezoidal, the triangular, and the sinusoidal to the added sinusoidal, there occurs an increase in the range of gradients. With the spreading of directions, the amount of bias in the estimated flow decreases, as shown in Figure 3.10, which explains the decrease in the perceived illusory motion found in the experiments.

Khang and Essock [38] have experimentally found the illusory motion to be strongest when the rectangular grid has elements with size 15-50 min in height 4-8 min in width — corresponding to a lower frequency component of at least 0.6 cpd and a higher frequency component of less than 7.5 cpd.

In the reduced Ouchi illusion [27, 28] the inset and surround regions are each sine-wave gratings of the same frequency, oriented in different directions (figure 3.5); solving for the pattern motion requires combining measurements from



(a) (b)

Figure 3.10: Expected length of optical flow and expected error in angle for the following three gradient distributions. (a) Rectangular checkerboard pattern with 16 vectors in major and 4 vectors in minor gradient directions. (b) 12 vectors in major direction, two vectors each at 10° and 20° to the left and right of major direction, 2 vectors in minor direction, (90° from the major) and 2 vectors at 10° to the left and right of the minor direction. (c) Approximation to distribution of gradients of the function $f(x, y) = \sin(x) \cdot \sin(5y)$. The optical flow is $(0, 1)$ and $\sigma_s = \sigma_t = 0.15$. With an increase in the spread of gradient directions, a decrease in the amount of bias occurs.

both the inset and the surround. However, the requirement that one never combine measurements across a motion discontinuity requires a decision of whether a set of one dimensional component measurements comes from one common pattern motion or several. The residual of the least squares solution provides an apt measure; in the absence of noise, the least squares solution of any set of component motions deriving from a single pattern motion will be zero. In the presence of noise, a relatively uniform distribution of component directions also gives a consistently small residual. The problem is that, with noise, the residual of the least squares solution for measurements from a limited set of orientations is indistinguishable from the residual of a solution when measurements are combined from entirely different pattern motions.

Experimental results find the inset to be segmented consistently for grating frequencies between 5 and 12 cpd [27, 28]. For grating speeds ranging from 0.2 to 2 cycles per degree (in this experiment the component speed changes with the relative angle of the inset and surround), the human visual system is sensitive to spatial frequencies between approximately 0.5 and 14 cycles per degree [36]. The receptive field size of cells in the visual pathway responding to gradients of different frequencies varies; the size of the receptive field changes the accuracy with which a boundary can be detected. Figure 3.8d shows the residual of a least squares solution for measurements collected in very small regions — the sudden spike in the residual magnitude when the region contains both orientations

indicates a likely motion boundary. When collecting measurements from a larger receptive field, as in figure 3.8 e, the position of the boundary is much less clear.

Figure 3.8(d,e) also shows that even in regions of a single direction, there is a non-zero residual, the ratio of this background residual to the residual calculated from measurements from differently oriented components determines whether or not the pattern is perceived as coherent. Figure 3.11 presents the ratio of the background residual to the residual of a receptive field on the boundary; all points are computed with a Monte Carlo simulation adding Gaussian noise to the derivative measurements, points marked with an "x" show component motion angles tested in the reduced ouchi stimulus (in [27], those with a "+" are component motion angles in experimentally tested symmetric plaid stimuli [39]. This ratio give a clear separation (on the y-axis) of the stimuli that were judged to be coherent (those whose orientation difference was less than or equal to 45°), and those that either gave a segmentation in the reduced ouchi stimulus or non-coherent perception of plaid motion (the rest).

For plaid patterns that are perceived as coherent, we can predict the bias in the perceived direction; Figure 3.12 explicitly considers the plaid patterns considered by Smith and Edgar [57]. Since for gratings of different spatial frequency 90 degrees apart, the bias is in the direction of the major gradient direction (motion direction of higher spatial frequency) the estimated flow of the plaid should be

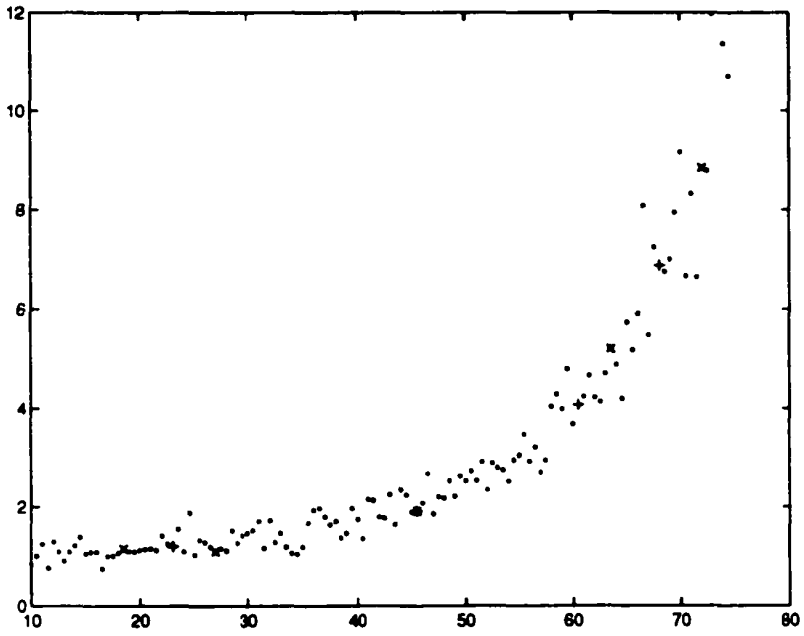
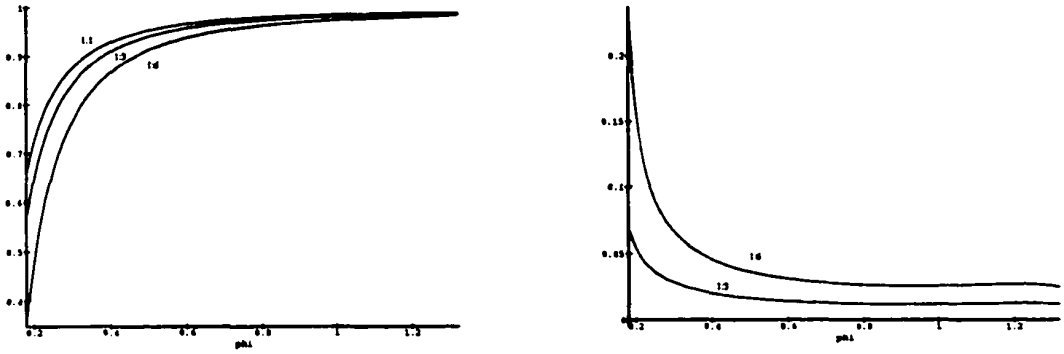


Figure 3.11: Ratio of background residual to boundary residual for noise $\sigma = 0.1$. The x-axis is one half of the angle between component motion directions.

closer in direction to the motion of the higher spatial frequency grating than predicted by the IOC model.

In summary, recalling that the bias in plaid velocity is in direction of the eigenvector corresponding to the largest eigenvalue of $\mathbf{M}^{-1}$, we can directly map changes in the plaid pattern to the expected bias. If the contrast of one component sine-wave grating increases, the major eigenvector moves towards the direction of motion of that component. For components of equal contrast and frequency, the major eigenvector is the vector average of the component motion vectors. In type II plaids, where both component motion vectors are on the same side of the IOC motion, this gives a bias towards this vector average di-



(a) (b)

Figure 3.12: Expected length of optical flow and expected error in angle for gradient distribution with measurements in the ratio (6 : 1, 3 : 1 and 1 : 1) at angle φ and $\pi - \varphi$ from the x axis. The actual flow is along the y axis and of length one. For the symmetric distribution in the ratio 1 : 1 no error in angle occurs.

rection. The effects of different frequencies can be modeled as different numbers of measurements in each direction; this also changes the direction of the major eigenvector. In the absence of noise, the residual of the least squares solution to the optic flow would be a perfect measure of the existence of a flow boundary — modeling the effects of noise on the magnitude of the residual proves that it is a measure of human perception of non-coherent motion.

3.6 Correcting the Bias?

In the statistical literature the model we used to describe the estimation of flow is referred to as the classic, "Errors-In-Variable" (EIV) model. It is usually

expressed in the notation $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = \mathbf{A_0} + \boldsymbol{\delta}\mathbf{A}$ and $\mathbf{b} = \mathbf{b_0} + \boldsymbol{\delta}\mathbf{b}$ where $\mathbf{A_0}$ and $\mathbf{b_0}$ are the true but unobservable variables (in our case the actual spatial and temporal derivatives $I_{x_i}, I_{y_i}, I_{t_i}$ at points i), $\boldsymbol{\delta}\mathbf{A}$ and $\boldsymbol{\delta}\mathbf{b}$ are the measurement errors, $\mathbf{A}$ and $\mathbf{b}$ are the corresponding observable variables and $\mathbf{x}$ the unknown parameters to be estimated (in our case u and v).

It is well known from the literature that estimation with least squares (LS) generally provides an inconsistent and biased estimate of the true parameter $\mathbf{x}$. The LS estimator gives an unbiased solution only for the regression model, that is, when $\boldsymbol{\delta}\mathbf{A}$ is considered to be zero and measurements $\boldsymbol{\delta}\mathbf{b}$ are independent, zero mean and equal distributed. The literature on estimation theory also provides a wealth of information on techniques dealing with the EIV model and how to compensate for the bias. However, to apply these techniques to the problem of flow estimation for navigating vision systems is computationally very difficult. In many situations, theoretically it should be possible to improve upon the estimation, but the particular stimuli discussed here pose problems for any statistical procedure.

The motion field on the eye of a moving system is due to the relative motion and the distance between the system and the scene in view. The assumption of constant flow is strictly true only for a scene consisting of a fronto-parallel plane moving with translation parallel to the image plane. To cope with general motions and scenes, the processing of flow has to be carried out in several stages.

In a first stage, normal flow measurements should only be combined very locally to generate an estimate of optical flow in a small patch of the image. In following stages, flow measurements of neighboring patches can be compared to find larger regions of common 3D motion or to delineate motion boundaries.

These considerations exclude models which assume that the motion component in each direction is computed over very large areas and then the single components are combined into a common 2D motion estimate with the simple IOC rule. Such models would not lead to biased optical flow estimates; however, they are of very limited applicability and cannot be applied to small areas. If flow is computed only within small image regions from few image measurements, it is statistically not justified to simply intersect the motion components in single directions and not to consider the amount of measurements involved. As a small number of measurements in one direction gives rise to a much larger variance in the noise (infinite variance for directions with a single measurement) than a larger number, this could lead to large errors in the flow estimate.

Any statistical technique to compensate for the bias requires knowledge of the statistics of the noise. For the noise model considered in the previous sections, this means knowledge of the covariance matrix of the noise vector $(\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_t)$. If such is available, the bias in the least squares estimation could be removed. If the model of constant flow is valid, this can be achieved with the "Corrected

Least Squares" estimator. If a more complicated model of general smooth flow within an image patch is necessary, iterative techniques have to be employed.

However, the major problem lies in the acquisition of the statistics of the noise. The noise parameters are not intrinsic to the system, but depend on the viewing situation and the scene in view, and in general the statistics can only be considered patch-wise constant. The noise parameters have to be estimated from the flow estimates within a spatiotemporal neighborhood by using the model which relates the image derivatives and noise to the flow estimates. However, from a limited amount of data, it is very difficult to obtain good estimates. Furthermore, the variance in the motion estimates turns out to be large with respect to the bias. For example, in simulations (see Figure 3.13), it has been found that for a noise level of 10% (that is, $\sigma_s = \sigma_t = 10\%$ of the value of the spatial gradient and the length of the flow) the standard deviation is twice as large as the bias. Thus, correction, even with an accurate estimate of the bias, in many cases would lead to a worsening of the solution.

In the particular situation of the Ouchi illusion, the 3D motion (either due to random eye movement or jiggling motion of the paper) changes rapidly. This makes the temporal integration of measurements very difficult as the system has only a short time-span to obtain the noise parameters.

In recent years the nonlinear estimator of "Total Least Squares" has received a lot of attention and it has also been applied to the problem of flow estimation

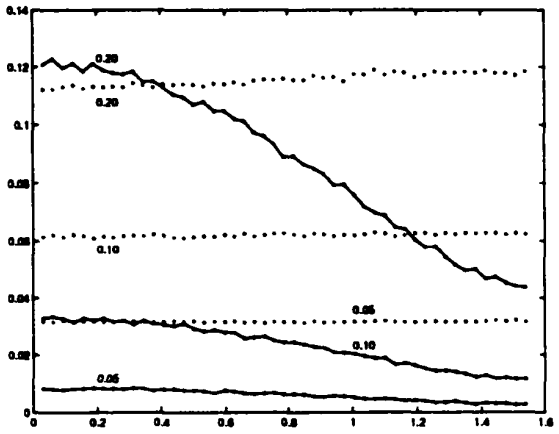


Figure 3.13: Expected error in value of length (solid lines) and standard deviation (dotted lines) obtained by a Monte Carlo simulation using Gaussian noise for three different standard deviations: $\sigma_s = \sigma_t = 0.2, 0.1$ and 0.05. The optical flow is $(0, 1)$, the magnitude of the spatial gradients is one, and gradients are distributed with 15 vectors in the direction at angle θ from the x axis and 5 vectors at angle $\pi/2 + \theta$.

[64, 65]. This estimator has been shown to provide an asymptotically unbiased solution for the EIV model in the case of white noise, that is, if the noise values are independent, and identically distributed. To whiten the noise, however, again it is necessary to obtain its covariance matrix. Without whitening, total least squares also gives biased solutions. In addition, total least squares is known to perform very poorly if outliers are present, and these are difficult to detect from few measurements.

The estimation and interpretation of optical flow from a statistical point of view has received attention before in the computational literature [12, 13, 24, 54, 61, 65]. Most relevant is the study of Nagel [45]. He considers an error model slightly more complicated than ours. In particular, he assumes Gaussian noise in

the gray values, linearly varying flow, and he considers the dependence of the partial derivatives in the gray value function at neighboring pixels arising from the computations with filters of discrete size. To compute the flow within neighborhoods he suggests an iterative estimation technique and then he uses hypothesis tests to compare neighboring flow estimates for whether they are compatible. In a subsequent study, Nagel and Haag [46] use this error model to find the error arising from least squares estimation and compensate for it for the purpose of tracking. The bias they find with their model is similar to our formulation. However, they only interpret this bias with regard to the underestimation in the length of the flow, and do not discuss the effects on its direction due to the distribution of the image gradients. Also, they do not discuss how to obtain the parameters of the noise distribution, but assume it to be available.

From a computational point of view, the problem of flow estimation is very difficult. In order to obtain very accurate flow estimates, a sufficiently large number of normal flow measurements is necessary. This means that data has to be spatially and temporally integrated via further computational models. As such computational models are based on assumptions about the 3D motion and the scene in view, they are not generally valid for systems moving in varied environments. The integration is possible only in image patches where the data is approximated well by the model. Thus, for the system to use a certain model, it first has to test its validity. For example, in order to employ a model of

smooth flow within a spatiotemporal neighborhood, the system has to check for discontinuities in a spatial neighborhood, verify that the flow doesn't change abruptly between frames and evaluate how well the flow is approximated by the particular model used. Clearly, these computations cannot be carried out on the basis of one-dimensional image velocity measurements alone, but require further spatiotemporal 3D information.

3.7 Conclusion

We have shown the problems of estimating two-dimensional image velocity from local one-dimensional motion measurements from a statistical point of view. As noise affects local motion measurements, that is, normal flow vectors, in both length and direction, the estimation of optical flow is biased. Theoretically, the design of any unbiased estimator would require knowledge of the statistics of the noise which often is hard to obtain. The only robust solution is to use local image measurements to directly constrain the camera motion and the scene structure. The following chapters lie within this framework.

Chapter 4

Constructing Models from One Viewpoint

A differential reconstruction is the representation of the scene structure which can be extracted just from the image derivatives from one viewpoint of a video sequence. In order to create useful models of scene structure, it is necessary to extend the basic constraints described in Section 2.4, relating the camera motion, the image derivatives, and the point depths. This chapter leads in two fundamental directions. First, Section 4.1 defines a method for segmenting the scene into regions that are well approximated by planar patches. Previous approaches either assume that such a segmentation is available, that the scene is globally smooth so that any segmentation is satisfactory, or try to post-process the scene to determine a segmentation. Section 4.2 gives examples of reconstructions of various scenes from a single viewpoint. Second, Section 4.3 provides experimental evidence that the standard error measures used to determine camera ego-motion are incapable of finding the correct direction of translation; there is a one-dimensional

subset — a valley — of motion parameters that all give low error. This is consistent with recent theoretical results, and is why we only use the differential image measurements only to define sets of plausible motion estimates. These sets are refined and explored in Chapter 5 when linking reconstructions from disparate viewpoints.

4.1 Segmentation

The assumption that each image region can be approximated by a plane requires a segmentation of the image into patches that avoid depth discontinuities. The segmented depth model has two parts: an assignment function $\mathcal{A}$, assigning image points to patches, and the positions in space of the planar patches, parameterized by $\vec{q}_1, \ldots, \vec{q}_n$. Together, these define the depth of every image point $\vec{r}_i$:

$$\frac{1}{Z_i} = \vec{q}_{\mathcal{A}(r_i)} \cdot \vec{r}_i$$

Finding the "best" set of depth planes for a given translational estimate requires finding the best camera rotation, the best assignment of pixels to patches, and the best position and orientation of those patches in space. This is equivalent to the general combinatorial problem of fitting a set of n planes and a parameter $\hat{\omega}$ to a cloud of points whose position is parameterized by $\hat{\omega}$.

It is not clear how the minimum of this function can be analytically found; this is a function of $3 \times n + 3$ parameters, with many local minima. We assume that we

have some estimate of the camera translation t, and estimate a solution with an iterative process. The iteration begins by defining random convex regions of the image and solving for the depth planes that best fit these image measurements. Then, iteratively re-assign image points to pieces of the depth model that best fit the local image measurements, and recompute the best fitting planes for the new assignments. More formally, the two steps are:

1. Solve the following linear system for $\vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_n$, and $\boldsymbol{\omega}$:

$$\forall_i : \; I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\vec{\mathbf{q}}_{\mathcal{A}(r_i)} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \mathbf{t} = 0$$

2. For each pixel, reset the assignment function.

$$\forall_i : \mathcal{A}(r_i) = argmin_j \left| V_i(u_n - \mathbf{u}_{\mathbf{rot}}(\hat{\boldsymbol{\omega}}) \cdot \mathbf{n}_i - (\vec{\mathbf{r}}_i \cdot \vec{\mathbf{q}}_j)(\mathbf{u}_{\mathbf{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}_i)) \right|$$

The initial assignment, the first reassignment, and subsequent iterations are show in Figure 4.1.

The segmentation converges rapidly to a good solution in a scene with several sharp depth discontinuities. For scenes with such properties, the iterative process typically converges in fewer than 20 iterations and gives a segmentation of the image into large patches that avoid depth discontinuities (Figure 4.2).

The convergence properties of this algorithm depend upon the scene in view and the camera motion. There are two problematic cases. When the camera translation is small, the image derivatives are mostly due to the camera rotation.

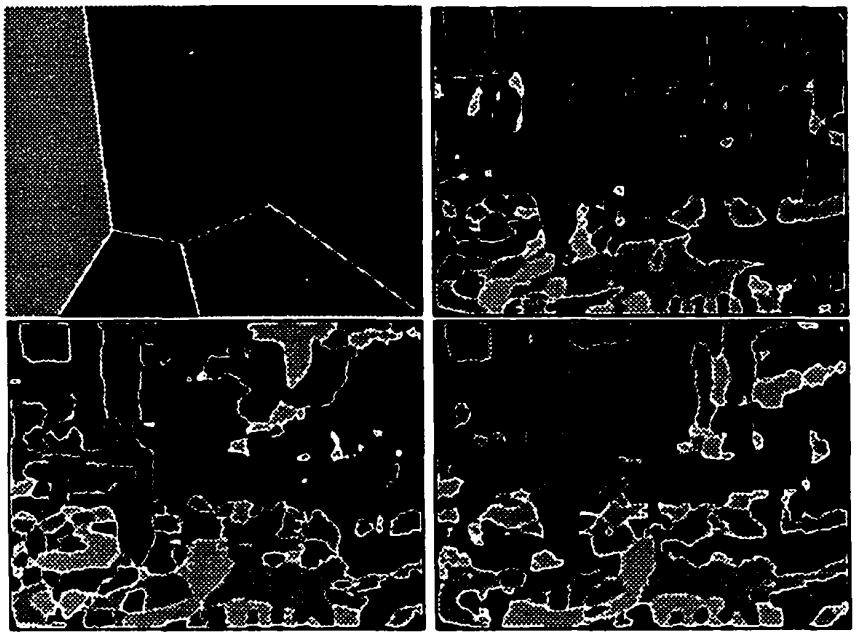


Figure 4.1: Discontinuity avoiding scene segmentation. An iterative process recursively modifies the scene structure model and defines image regions consisting of points whose scene depths are related. The assignment at the initial condition, and after the first, fifth and fourteenth iteration.

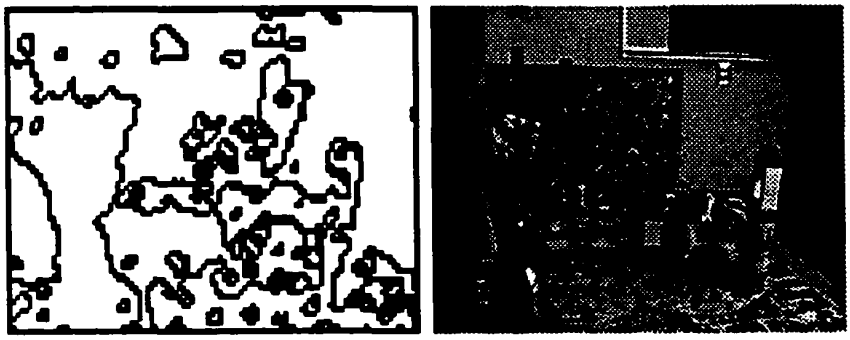


Figure 4.2: For scenes with sharp discontinuities, the iterative algorithm results in segmentation into large regions.

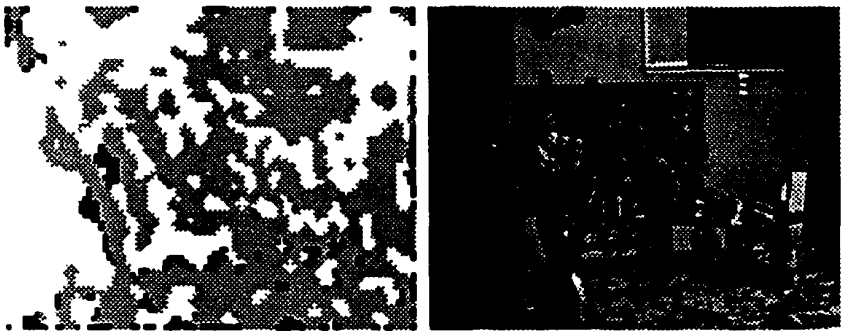


Figure 4.3: The segmentation process depends both upon the scene, and the camera motion. When the camera translational velocity is small, the segmentation may not converge.

Then, the noise in these measurements swamps the remaining information relating to relative depths. This leads to an arbitrary, noisy, segmentation, where unrelated scene points which are far apart may be assigned to lie on the same plane. Figure 4.3 shows the segmentation algorithm for the same scene, from a part of the video sequence with a much smaller translation velocity. The second problem occurs when the scene is mostly smooth but non-planar. In this case, there are many ways of approximating the scene with planar patches, each method approximates some regions better than others, and the algorithm does not converge because there is not a clearly optimal assignment. One example of this sort are shown in figure 4.4.

In both of these problematic cases the segmentation fails because it is implicitly searching for a piecewise planar representation of the surface, with a few large planar patches. When the scene is view is not well approximated by a few, large, planar patches, or when the necessary information is subsumed by noise,

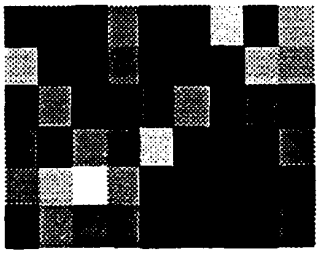
Figure 4.4: Because a small number of large planes does not approximate large round surfaces, the segmentation may not converge.

it is necessary to change the target representation. Instead of seeking a segmentation that corresponds to objects in the scene, it is possible to more robustly define a larger set of much smaller patches, enforcing the condition that patches never cross depth discontinuities (but perhaps there are many patches on a single continuous object). The straightforward algorithm is pictorially illustrated in Figure 4.5.

This segmentation can be used in a feedback loop, to avoid problems when using normal flow measurements derived from image derivatives whose filter support crosses discontinuities. This works best if the segmentation has created large image regions; then the feedback system erodes each patch to find a central region such that every image pixel used by the derivative filter is entirely within the patch. These eroded patches are pictured in Figure 4.6.

For each pixel i, let x_i be the patch label when the scene is segmented using previous algorithm.

For each pixel i, let y_i be the patch label when segmented into small square blocks .



Group together all pixels with the same pair (x_i, y_i)

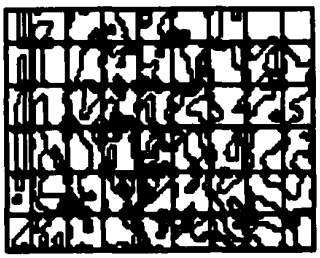


Figure 4.5: When it is not possible to create a segmentation with only few elements, it may still be possible to make a set of smaller image patches which respect discontinuities in the scene.

4.2 Differential Reconstructions

As in [7], creating a differential reconstruction requires finding the best translation $\mathbf{t}$. This solution proceeds as a search in the space of translational directions. After the segmentation process, we consider $\mathcal{A}$ to be fixed, so the error function is dependent only on $\mathbf{t}$. The exact formulation of the error function depends on the choice of scene representation. In this section we consider a set of arbitrarily oriented planes, and a piecewise planar, continuous mesh. When creating the differential reconstruction it is not usually possible to robustly solve for higher order piecewise polynomial models of scene depth, although these constraints can be written in the same form. Equation 2.10 gave the constraints for a single polynomial scene model. Creating a set of independent polynomial surface patches or a spline representation involves the same segmentation process as arbitrary planar patches or the mesh representation. These higher order models may be appropriate in creating models combining data from disparate viewpoints.

4.2.1 Patch Reconstructions

The representation used in the segmentation process models the scene as a set of arbitrarily oriented planes. Each translational estimate defines the following over-constrained linear system, with unknowns $\boldsymbol{\omega}, \vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_n$.

$$\forall_i : \; I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\vec{\mathbf{q}}_{\mathcal{A}(\mathbf{r}_i)} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \hat{\mathbf{t}} = 0 \tag{4.1}$$

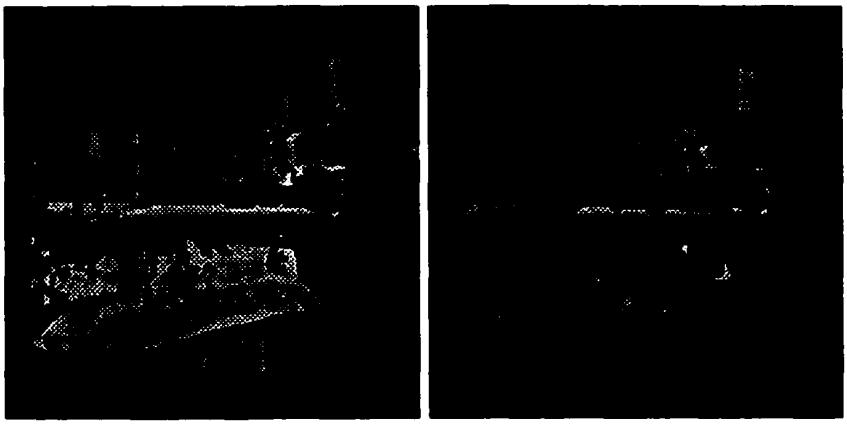


Figure 4.6: A new, segmented view of the paper bar(left), and, (right) the image regions believed to not include depth boundaries. On this data one can compute highly accurate motion estimates.

Let $f(\hat{\mathbf{t}})$ be the residual of the least squares solution to the system defined with translational estimate $\hat{\mathbf{t}}$.

$$f(\hat{\mathbf{t}}) = min_{\{\boldsymbol{\omega},\vec{\mathbf{q}}_1,\ldots,\vec{\mathbf{q}}_n\}} \sum_i (I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\vec{\mathbf{q}}_{\mathcal{A}(r_i)} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \hat{\mathbf{t}})^2 \qquad (4.2)$$

Some direction t minimizes the function f. Solving the linear system defined by that translation t gives values for $\vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_n$, and $\hat{\boldsymbol{\omega}}$. This is the differential reconstruction available from one set of image measurements. It defines the position and orientation of each patch, and the instantaneous translational and rotational velocity of the camera. One such reconstruction is shown in Figure 4.6. This reconstruction uses the segmentation whose iterative steps and final assignment where illustrated earlier in Figure 4.1.

4.2.2 Mesh Reconstructions

Another representation is a piecewise planar mesh, made up of patches that must meet their neighbors and have no discontinuities. This requires a triangular subdivision of the image plane and a different method of computing the $\frac{1}{Z_i}$ value at each pixel. Let $p_1, \ldots, p_k$ be the vertex set of the triangular subdivision. The (unknown) scene depth of each of these points is $q_1, \ldots, q_k$. (Here we maintain the use of the variable q as the "unknown scene depth variable", even though now it is the depth of a mesh vertex instead of the normal to a planar patch). Then, the scene depth of an image point $\mathbf{r}_i$ which is inside triangle p_a, p_b, p_c, can be expressed as:

$$\frac{1}{Z_i} = \alpha_a q_a + \alpha_b q_b + \alpha_c q_c \tag{4.3}$$

where $(\alpha_a, \alpha_b, \alpha_c)$ are the barycentric coordinates of a r_i in triangle p_a, p_b, p_c. If we create the vector $\vec{\lambda_i}$ with three non-zero elements for each point r_i so that

$$\vec{\lambda_i} = \langle \ldots, 0, \ldots, \alpha_a, \ldots, 0, \ldots, \alpha_b, \ldots, 0, \ldots, \alpha_c, \ldots, 0, \ldots \rangle \tag{4.4}$$

then we can write the scene depth of each point as a linear function of the depth of all the mesh vertices: $\frac{1}{Z_i} = \vec{\lambda_i} \cdot \langle q_1, \ldots, q_k \rangle$. The linear system constraining all the parameters of the differential reconstruction has the familiar form:

$$\forall_i : \; I_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega} + (\langle q_1, \ldots, q_k \rangle \cdot \lambda(\mathbf{r}_i))\vec{\mathbf{a}}_i \cdot \mathbf{t} = 0 \tag{4.5}$$

This representation is suitable for scenes without sharp depth discontinuities. One such scene is the rigid part of the standard Yosemite test sequence. Figure 4.7

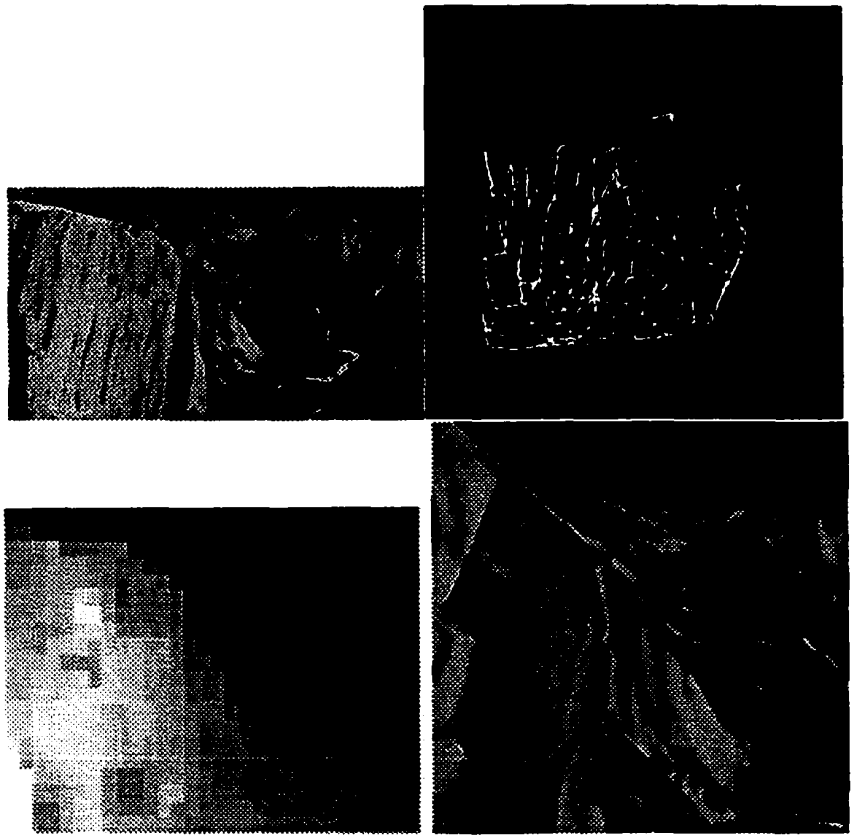


Figure 4.7: An image of the standard Yosemite test sequence, the computed mesh structure, and a reprojection of the scene to a new viewpoint.

shows one image of this sequence, the mesh structure, a depth map, and the reconstruction of the scene projected to a new viewpoint.

4.3 Experimental

Given a segmentation that respects depth boundaries in the scene, we can study the topology of the error surface. The error surface is the error defined for every possible translational direction. We can ask the question: "How well do

different translations fit the observed image derivatives". Since we can express the constraint as a linear system for a given translation, it is possible to try a great deal of translations rapidly. This allows the exploration of the topology of this error surface for a variety of scenes.

As an initial example Figure 4.8 illustrates the error surface for the Yosemite scene. The figure shows the field of view of the camera, and the error surface. The known (ground truth) translation is shown as an arrow piercing a sphere around the camera center. This arrow is the direction in which the camera is traveling. The sphere is a way of representing all the possible translational directions. For each direction $\hat{\mathbf{t}}$, the error of $f(\hat{\mathbf{t}})$, from Equation 4.2, is encoded as a gray value; white is smaller error. Only a cut out of the sphere is shown. The minimum of this error function is well defined, can be accurately localized, and fits the known ground truth. Because the camera motion is mostly towards the scene (as opposed to sideways), theoretical results suggest that there is little ambiguity in such a case.

The topography of the error surface varies qualitatively for different scenes. For scenes where the objects are far relative to the camera motion, there is little information, because the translational component of the image motion at a point on the image is proportional to inverse of the distance to that point. Therefore these scenes are the most likely to have ambiguous motion.

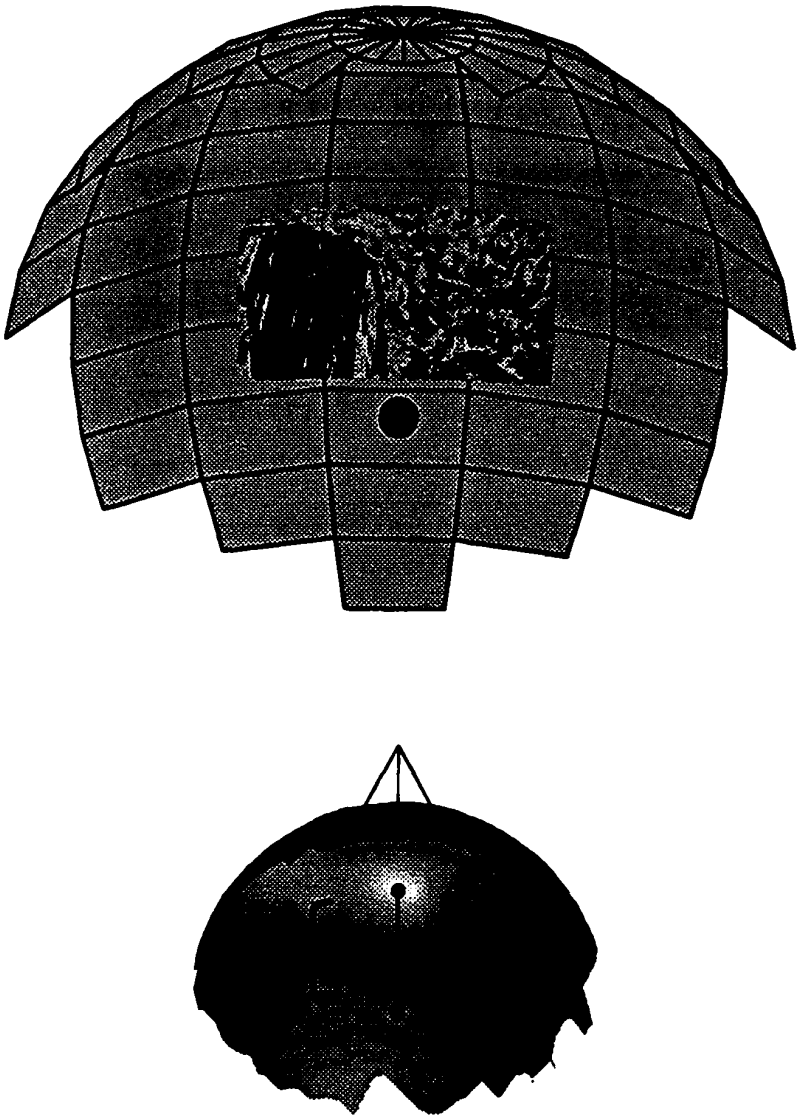


Figure 4.8: On top is an illustration of the camera geometry, with the camera nodal point represented by the dark dot. The image covers only a small part of the space around the camera, but since the true translation (arrow) passes close to the image, the error surface has a clear minima (bottom).

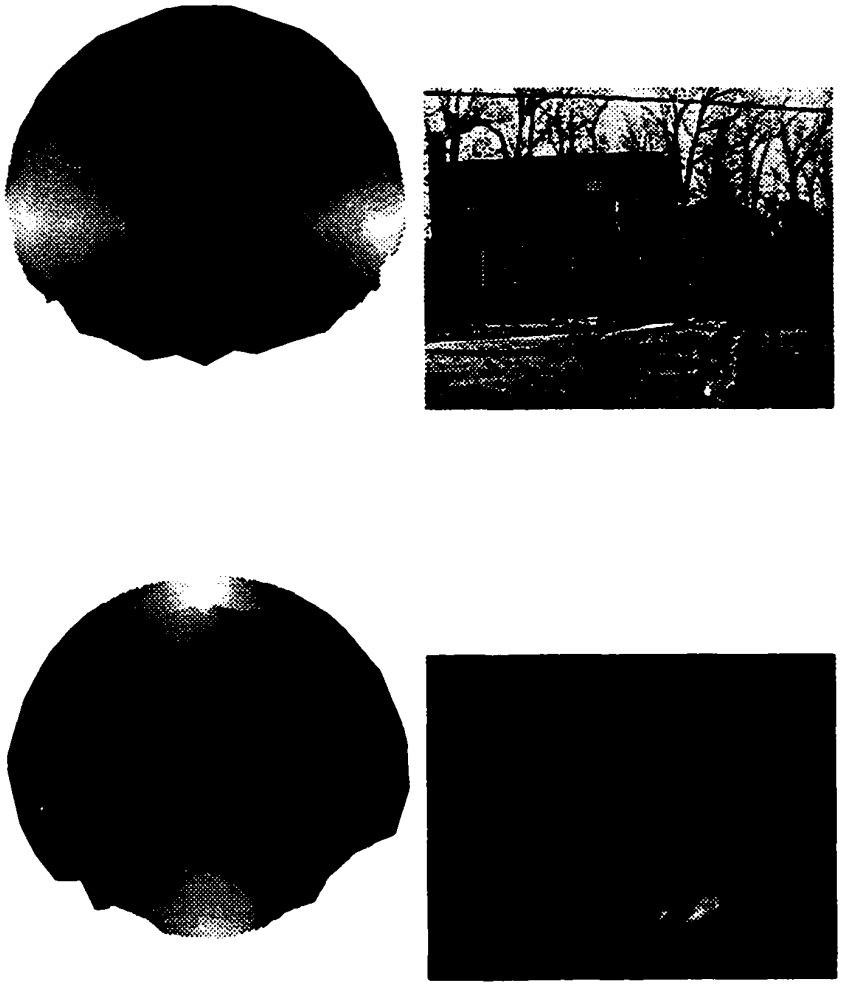


Figure 4.9: Outdoor scenes, where the camera motion is small relative to the distance to objects in the world. The second scene is looking at clouds, effectively a plane at infinity.

Figure 4.9 shows the error function for two different outdoor scenes, both far from the camera. In both cases, the error function has a large region of translational directions giving low error. These directions fall in a broad swath that is oriented towards the image center. The second case is particularly interesting, the camera is looking vertically, towards a collection of clouds, a scene that is effectively approximated by a plane at infinity. However, if the $\frac{1}{Z}$ term is zero for every pixel, from Equation 2.8 there should be no constraint at all on the t. The answer is rather prosaic; this was a windy day, so the clouds had a rather large absolute velocity. This picture is then an illustration of the well known ambiguity confounding rotation and translation for a camera with a small field of view.

Close scenes that have simple structure may also be problematic. There exists a set of surfaces for which even a zero noise motion field may be ambiguous. That is, there exists camera motions and surfaces such that, a different camera motion, and a different surface can lead to exactly the same image measurements. These two interpretations cannot be distinguished. These surfaces are ruled conics [30]. This is a small set of surfaces, and the ambiguity is only present for particular positions of the camera relative to the surface, but any scene where the depth can be reasonably approximated by such a conic is likely to be ambiguous in the presence of noise. Figure 4.10 gives examples of scenes with relatively simple structure that are close to the camera.

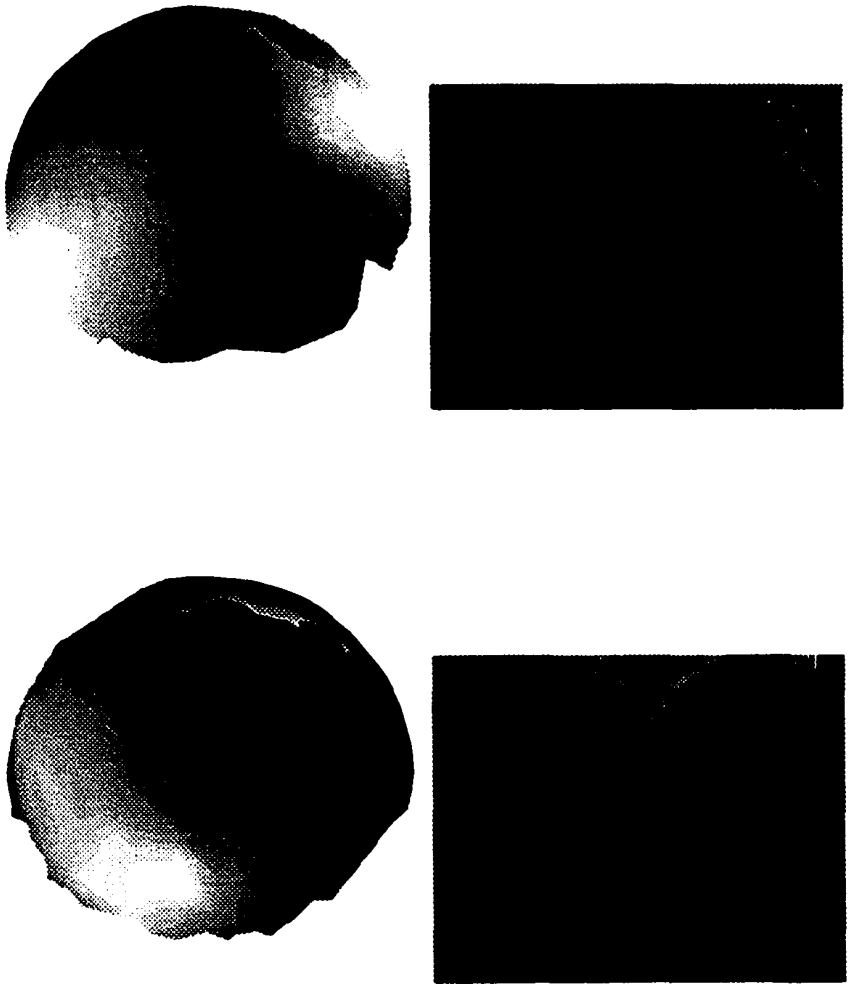


Figure 4.10: Outdoor scenes, (top) a view down over a step in an amphitheater, and (bottom) sideways view along a sidewalk. The camera motion is large relative to the distance to objects in the world. For these scenes, the ambiguity may disappear for some real camera translational directions.

Figure 4.11: The error surface for the scene shown in Figure 4.2. When the translation is large and towards the image, the error function has a clear minimum. In other cases, there remains an ambiguity.

A scene with sharp depth discontinuities, and a relatively large translation, gives the best constraints on the camera motion 4.11. The motion is strictly constrained — either to a particular point, when the camera is moving forward, or along a valley. This valley passes through the correct translational direction and is oriented so that it would pass through the image center if it were extended.

4.4 Reconstruction from Motion Valleys

The set of possible translations that minimize Equation 4.2 often lie along a valley instead of around a point. Figure 4.12 shows the shape reconstructed from different translations along this valley. The motions along the minimal valleys all give plausible depth maps. What is important in this set of reconstructions is their qualitative similarity. None of them have a rugged structure, and each reconstruction, along with its estimated translation, predicts almost the same set of image measurements. Therefore creating the full reconstruction does not, by itself, give additional information to constrain the camera motion.

4.5 Conclusions

From one viewpoint it is theoretically and practically very difficult to compute accurate camera motion. In general, one view robustly defines an approximately one dimensional set of "plausible motions", the set of translational directions in

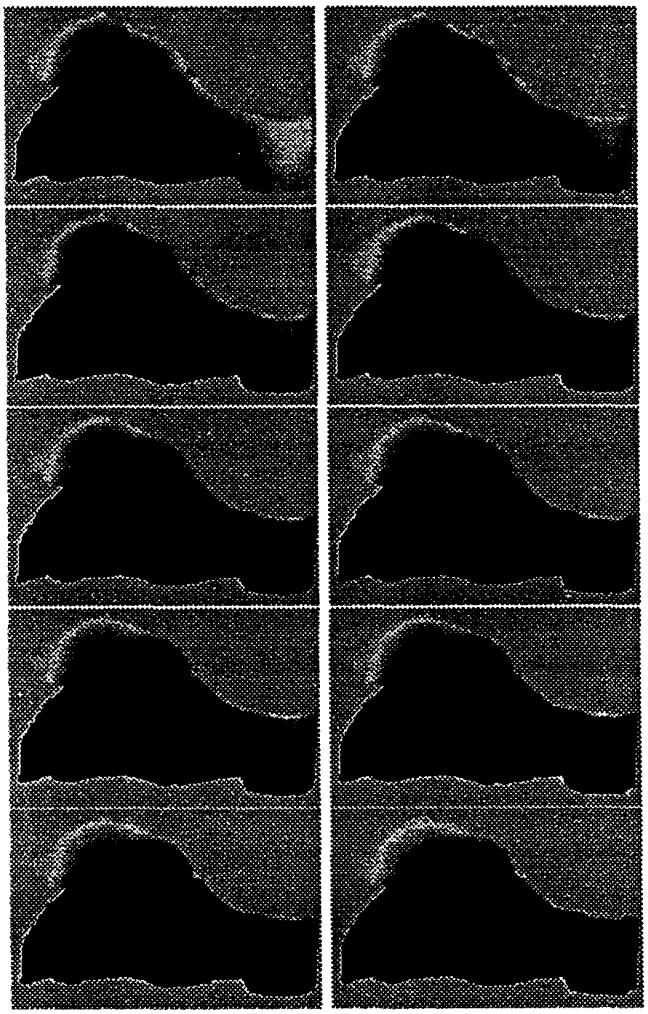


Figure 4.12: **Reconstructed depth maps for different translations along the valley minimum.**

the valley of the error function. Each translation has a corresponding rotation and depth map which approximately fit the image data. In some cases this depth map has regions of significant negative depth, which can be ruled out a-priori. However, there typically remains a set of camera translation directions which all approximately minimize Equation 4.2. Further constraints on the motions are only possible given new information. The next chapter discusses linking the coordinate systems of reconstructions from different viewpoints. The new viewpoint gives the additional information necessary to accurately find the camera motion.

Chapter 5

Linking

To obtain more accurate, robust solutions, it is necessary to combine visual information from camera viewpoints that are far apart. Chapter 4 defines a method for creating a 3D reconstruction from one viewpoint. Because we assume that the scene stays constant, the reconstructions made from two viewpoints have a nice relationship, as long as each reconstruction is created using the correct motion parameters. This relationship is simple to define in the ideal case. In the general case, searching for this reconstruction amounts to a 2D search — specifically, searching along the one dimensional set of possible translation directions defined by each viewpoint

5.1 Differential Linking

The first instance of coordinate system linking concerns two motion fields, captured from viewpoints that are very close to each other — as will be the case

for sets of image derivatives whose temporal derivative filters are centered on subsequent frames of a video. We assume that we already have a camera motion estimate $(\hat{\mathbf{t}}^{(1)}, \hat{\boldsymbol{\omega}}^{(1)})$ for the first motion field, and a camera motion estimate $(\hat{\mathbf{t}}^{(2)}, \hat{\boldsymbol{\omega}}^{(2)})$ for the second motion field. The problem is that each of the translation vectors has an arbitrary scale factor; within each differential reconstruction one can assume that the translation has unit magnitude which defines a scale for the reconstruction. When combining two differential reconstructions, it is necessary to find the scale factor relating them.

The translation in the initial frame $\hat{\mathbf{t}}^{(1)}$ can still be set to an arbitrary magnitude. We define the scale of our 3D reconstruction so that $\hat{\mathbf{t}}^{(1)}_{3D} = \hat{\mathbf{t}}^{(1)}$. Since the scene is assumed to be constant, so should the reconstruction from frame to frame be constant. Finding the correct scale for $\hat{\mathbf{t}}^{(2)}$, amounts to finding the magnitude of the 3D translation vector $\hat{\mathbf{t}}^{(2)}_{3D}$ which creates a scene reconstruction at the same scale. This requires the solution to linear system, defined by a constraint at every image pixel in each flow field:

$$k_1(I^{(1)}_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega}^{(1)}) + (\vec{\mathbf{q}}_{\mathcal{A}(\mathbf{r}_i)} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \mathbf{t}^{(1)} = 0 \tag{5.1}$$

$$k_2(I^{(2)}_{t_i} + \vec{\mathbf{b}}_i \cdot \boldsymbol{\omega}^{(2)}) + (\vec{\mathbf{q}}_{\mathcal{A}(\mathbf{r}_i)} \cdot \mathbf{r}_i)\vec{\mathbf{a}}_i \cdot \mathbf{t}^{(2)} = 0 \tag{5.2}$$

Then, setting $\hat{\mathbf{t}}_{3D_2} = \frac{k_1}{k_2}\hat{\mathbf{t}}_2$ gives the translation motion vector for the second motion field. Solving for both k_1 and k_2 is more stable that solving for a single scale factor. This is equivalent to the more intuitive system to find the scale factor, where k_1 multiplies directly the $\mathbf{t}(1)$ estimate. This system also has the

advantage that it is linear in k_1, k_2, and $\vec{\mathbf{q}}_{\mathcal{A}(r_i)}$. Thus the solution gives both the relative scale factor and and scene depth model using image measurements from two seperate frames. However, informal experiments have shown that this does not appreciably improve the differential reconstructions.

A scene reconstruction using the second motion field, and the translation $\hat{\mathbf{t}}_{3D_2}$ will have the same scale factor as the reconstruction created from the first motion field using the translational estimate $\hat{\mathbf{t}}^{(1)}$. This constraint can be "chained", to put many differential motion estimates into a common coordinate system.

Given the differential motions $(\hat{\mathbf{t}}^{(1)}, \hat{\boldsymbol{\omega}}^{(1)}), (\hat{\mathbf{t}}^{(2)}, \hat{\boldsymbol{\omega}}^{(2)}), \ldots (\hat{\mathbf{t}}^{(i)}, \hat{\boldsymbol{\omega}}^{(i)})$, the discrete rotation $\mathbf{R}_i$ which transforms directions in the ith frame into directions in the coordinate system of the first frame is:

$$\mathbf{R}_i = B_1 B_2 \ldots B_{i-1} \tag{5.3}$$

B_i is the rotation matrix which corresponds to the displacement caused by the angular velocity $\boldsymbol{\omega}_i$ over one frame.

$$B_i = (I - \frac{1}{2}[\boldsymbol{\omega}]_\times)^{-1}(I + \frac{1}{2}[\boldsymbol{\omega}]_\times) \tag{5.4}$$

where $[\boldsymbol{\omega}]_\times$ is a skew-symmetric matrix corresponding to the cross product with vector $\boldsymbol{\omega} = [\alpha, \beta, \gamma]^{\mathrm{T}}$:

$$[\boldsymbol{\omega}]_\times = \begin{pmatrix} 0 & -\gamma & \beta \\ \gamma & 0 & -\alpha \\ -\beta & \alpha & 0 \end{pmatrix} \tag{5.5}$$

The other component of the rigid transformation between from frame i to frame 1 is the translation T_i. This must be calculated recursively, at the same time as the scale factor (because we can only find the scale factor between subsequent frames):

$$\begin{aligned} s^{(1)} &= 1 \\ T^{(1)} &= [0,0,0] \\ s^{(i)} &= s^{(i-1)} * \frac{\hat{\mathbf{t}}_{3D_i}}{\hat{\mathbf{t}}_{3D_{i-1}}} \\ T^{(i)} &= T^{(i-1)} + s^{(i)}\mathbf{R}^{(i)}\hat{\mathbf{t}}^{(i)} \end{aligned} \tag{5.6}$$

One example of this frame linking is shown in Figure 5.1, where the differential motions were computed between consecutive frames for a 300 frame video sequence. These motions were linked into the coordinate system of a one viewpoint and are displayed together with the segmented differential reconstruction from that viewpoint. This frame linking is possible only because the camera motion direction was forward. Because the translation direction was within the field of view, the motion parameters are better constrained and not subject to sever ambiguities.

There are three limitations of this approach of linking frames. First, Equation 5.2 requires that the motion of the camera between the first and second motion field is small; specifically, the same segmentation must be valid for both motion fields, and the depth of each particular pixel must change minimally.

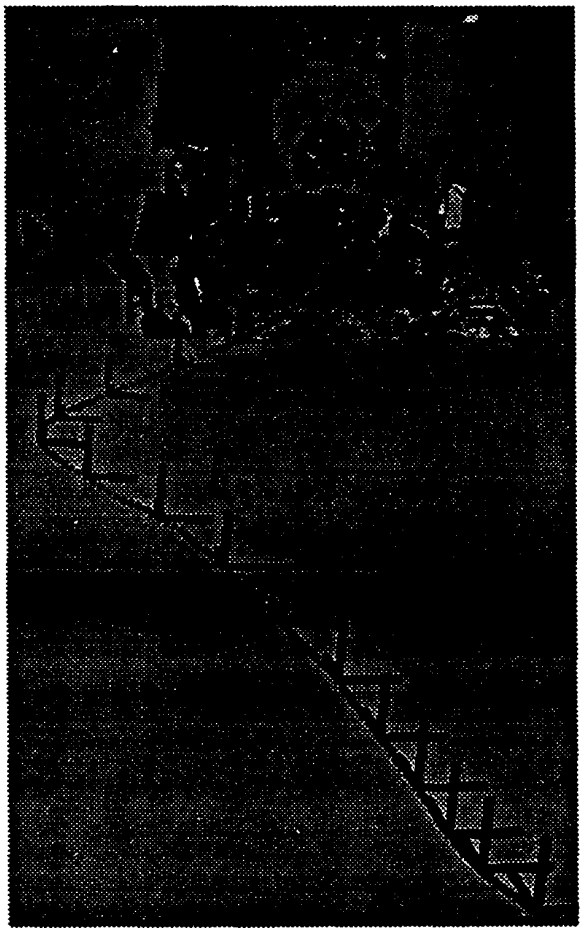


Figure 5.1: A differential scene reconstruction from a single normal flow field. The 3D motion of the camera was computed between consecutive frames and the positions plotted in the coordinate system of the reconstruction. A coordinate axis is shown every tenth frame to indicate camera rotations. The camera trajectory becomes erroneous after a camera jerk breaks the differential motion assumption.

These are reasonable assumptions in video sequences with a high frame rate. There may be statistical bias introduced in particular cases, such as when most surface patches are consistently oblique with respect to the camera optical axis. Second, because the scale factor is linked only between consecutive frames, there is no error correction. Small errors in computing the relative scale factor between frame one and two will be propagated throughout the entire set of frames to be linked. Third, motion fields from consecutive frames are unlikely to be able to resolve the motion ambiguities addressed in Section 4.3, because the camera motion is not likely to change sufficiently between subsequent motion fields.

The subject of the next section is to address all of these considerations by explicitly linking reconstructions from viewpoints that are further apart. The disparate viewpoints provide stronger constraints on the scene structure, but the constraints are no longer expressible as linear systems.

5.2 Linking Disparate Viewpoints

Looking at a scene from viewpoints that are far apart gives more constraints on the scene structure. More information is available than exists in motion fields from one viewpoint. Here we consider the case where we have captured motion fields from two different viewpoints. These motion fields are assumed to be somewhat overlapping, so that some of the scene is imaged in each viewpoint. Since the scene is assumed to be constant, the reconstructions from each viewpoint should

also match. We discuss methods to use the new constraints to find the correct differential camera motion. These methods build on top of the processing done on each motion field individually — that is, they assume that we have a set of candidate motions (and therefore candidate reconstructions) created separately from each viewpoint.

The methods and heuristics discussed in this section involve more computation than methods discussed in Section 4.3. It is also the case that there is, in fact, more time available between the capture of far apart viewpoints. The goal of the reconstruction is no longer a moment-to-moment approximation of the local scene, but rather a fully three dimensional scene model. We should not need to reconstruct such a complicated scene model in every frame, so the time constraints for a "real-time reconstruction system" are on the order of seconds instead of hundredths of seconds.

First, we need to introduce a modicum of new notation. In the coordinate system defined at the first viewpoint, let $V^{(1)}$ be the set (a valley) of candidate translations, $\mathbf{t}^{(1)}$ be the (unknown) correct camera translation, and $\hat{\mathbf{t}}^{(1)}$ be any estimated translation. We will only consider $\hat{\mathbf{t}}^{(1)} \in V^{(1)}$, the preprocessing of each motion field has reduced the two dimensional space of possible translation directions to this set $V^{(1)}$ of candidates. The same applies to the second viewpoint, $V^{(2)}, \mathbf{t}^{(2)}, \hat{\mathbf{t}}^{(2)}$, will be the set of plausible translational directions, the correct translation, and single candidate translations, respectively. Finally, there

is a rigid motion between the first viewpoint and the second viewpoint, defined by the rotation matrix $R^{(1\rightarrow 2)}$and the translation $T^{(1\rightarrow 2)}$.

In general terms, the algorithm to link disparate viewpoints is then the following:

1. Find $V^{(1)}, V^{(2)}$, the set of plausible translational directions from each frame.
2. For all pairs $(\hat{\mathbf{t}}^{(1)}, \hat{\mathbf{t}}^{(2)})$, with $\hat{\mathbf{t}}^{(1)} \in V^{(1)}$, and $\hat{\mathbf{t}}^{(2)} \in V^{(2)}$
   - Create the differential reconstructions using $\hat{\mathbf{t}}^{(1)}$ and $\hat{\mathbf{t}}^{(2)}$
   - With respect to some error function, solve for the best fitting equiform transformation that matches these reconstructions.
   - Record error measure for how well this pair of reconstructions can be matched.
3. Choose the pair $(\hat{\mathbf{t}}^{(1)}, \hat{\mathbf{t}}^{(2)})$ for which the reconstructions matched the best.

The remaining work is to define an error measure that indicates whether or not two reconstructions match. The following section considers the ideal case, an introduction to finding the equiform transformation between two reconstructions if corresponding points in each reconstruction are given.

5.2.1 The Ideal Case

Assuming zero noise in the image derivative measurements, and correct egomotion estimates, the 3D reconstruction from each viewpoint will be an accurate

Euclidean reconstruction scaled by an unknown factor. The relationship between two such (ideal) reconstructions is an equiform transformation, which is a rotation, a translation, and a uniform scaling. This corresponds to the rigid transformation between the viewpoints, and a resolution of the different scale factors used in each reconstruction.

The relationship between a point in the first reconstruction, $P^{(1)}$ and the same point in the second reconstruction $P^{(2)}$ is:

$$sR^{(1\rightarrow 2)}P^{(1)} + T^{(1\rightarrow 2)} = P^{(2)} \tag{5.7}$$

Three non-linear corresponding points suffice to completely constrain $R^{(1\rightarrow 2)}$and $T^{(1\rightarrow 2)}$. However since $R^{(1\rightarrow 2)} \in SE(3)$, the elements are not independent, and this cannot be expressed and solved as a linear system. If the coordinates of the 3D point were specified accurately, this constrained optimization problem can be formulated as a search for the minimum of the following function, and efficiently solved with a BFGS-Quasi Newton method:

$$\{R^{(1\rightarrow 2)}, T^{(1\rightarrow 2)}, s\} = argmin_{\{R^{(1\rightarrow 2)}, T^{(1\rightarrow 2)}, s\}} \sum_i \left| sR^{(1\rightarrow 2)}P_i^{(1)} + T^{(1\rightarrow 2)} - P_i^{(2)} \right|. \tag{5.8}$$

With real data, however, this method fails to take into account how 3D points are reconstructed. Creating the 3D coordinates of a reconstructed point P_i combines the image coordinates p_i and the computed depth at that point. Because of the inverse relationship between the depth of a point and its motion on the image, the differential reconstruction is actually computing the inverse depth at

a pixel $\frac{1}{Z_i}$. So small errors in image measurements will lead to small errors in the estimation of $\frac{1}{Z_i}$, but could lead to large errors in the absolute depth, or in position of reconstructed 3D points.

In order to alleviate this problem, we define a transformation between a point P in 3D, and a point $\zeta(P)$ in "reconstructed space". Specifically, if $P = (X, Y, Z)$, then $\zeta(P) = (\frac{X}{Z}, \frac{Y}{Z}, \gamma\frac{1}{Z})$, which are the image coordinates of the point augmented with the some linearly scaled inverse depth. The γ term is necessary — the image points and the inverse depth are computed with different processes so it is not clear what the appropriate relative weighting should be in measuring their error. Determining γ for one method of finding correspondence between different viewpoints is the subject of Section 5.2.2.

Since we are using standard non-linear minimization techniques to find the best equiform transformation between viewpoints, it is simple to now write the correct error function (where $\zeta(P)$ is transforms a point to "reconstruction space", and there is a parameter γ which does not appear):

$$\{R^{(1\to 2)}, T^{(1\to 2)}, s\} =$$

$$argmin_{\{R^{(1\to 2)}, T^{(1\to 2)}, s\}} \sum_i \left| \zeta(sR^{(1\to 2)}P_i^{(1)} + T^{(1\to 2)}) - \zeta(P_i^{(2)}) \right| \qquad (5.9)$$

The actual minimization of this function uses Rodrigues parameters to encode the rotation matrix $R^{(1\to 2)}$ with three independent parameters. This is the minimization function used in the remainder of this chapter.

5.2.2 Approximate Correspondence

The straightforward method of finding the equiform transformation between two viewpoints requires known correspondences between points visible in each viewpoint. The following algorithm generates a set of such correspondences between two disparate viewpoints from the same video sequence.

1. Define R_1 to be a set of k points in image 1.

2. For each frame: $j = 1 \longrightarrow n$

   (a) Compute $(\mathbf{t}^{(j)}, \boldsymbol{\omega}^{(j)})$ the camera ego-motion for frame j.

   (b) For each point $\mathbf{r}_i \in R_j$.

      i. Compute $\frac{1}{Z_i}$ from small patch around point $\mathbf{r}_i$

      ii. $\vec{\delta_i} = -\frac{1}{Z_i}(\hat{\mathbf{z}} \times (\mathbf{t}^{(j)} \times \mathbf{r}_i)) + (\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega}^{(j)} \times \mathbf{r}_i)))$

      iii. $\mathbf{r}_{i+1} = \mathbf{r}_i + \vec{\delta_i}$

3. The point sets R_1, R_n are approximately corresponding point sets.

This is an intentionally primitive algorithm; it ignores local image texture information. However, it possible to estimate the accumulated error as these points are tracked between frames. Let $\vec{\epsilon_i}$ be the error in propagating a point location between frame j and $j+1$ for some point i, assumed to be the same point for the remainder of this discussion. Suppose $\frac{1}{Z_i}$ has an error ϵ_i, and in general, $\frac{1}{Z}$ has an error that can be expressed as a zero-mean random variable

with variance ϵ. Then from Equation 2.4, the magnitude of $\vec{e_i}$ is bounded by ϵ_i (by definition, $\hat{\mathbf{z}}$, $\mathbf{r}_i$, and $\mathbf{t}$ are unit vectors):

$$|\vec{e_i}| = |\vec{\delta_i}^{\text{TRUE}} - \vec{\delta_i}| = |\epsilon_i(\hat{\mathbf{z}} \times (\mathbf{t}_i \times \mathbf{r}_i))| \leq \epsilon_i \tag{5.10}$$

This error is never fixed and it accumulates from frame to frame, leading to a total a final error in correspondence:

$$\vec{e}_{1 \to n} = \sum_{i}^{n-1} \vec{e_i} \tag{5.11}$$

Since the errors have some random component, we can do better than the trivial bound $|\vec{e}_{i \to n}| \leq n\epsilon$. Since $\vec{e_i}$ is dependent on $\mathbf{t}_i$, the accumulated error for the sequence depends on the consistency of the camera motion. Assuming the the errors in subsequent frames are independent, then if $\mathbf{t}$ is roughly constant for the sequence, the tracking error for each frame is perpendicular to to direction $\mathbf{t} \times \mathbf{r}$. What remains is to determine the the expected value of the magnitude of this error:

$$E(|\vec{e}_{i \to n}|^2) \leq E(\sum_{i=1}^{n}(\epsilon_i(\mathbf{t} \times \mathbf{r})))^2 \leq E(\sum_{i=1}^{n}\epsilon_i)^2 = \sum_{i=1}^{n} E(\epsilon_i^2) = n(\epsilon^2) \tag{5.12}$$

so,

$$E(|\vec{e}_{i \to n}|) \leq \epsilon\sqrt{n} \tag{5.13}$$

In other words, the expected tracking error over n frames is bounded by $\sqrt{n}$ times the error in estimating $\frac{1}{Z}$ in each frame. This simple form derives from the fact

that we use a normalized camera and we scale the translational motion in each frame to be of unit magnitude.

What is a reasonable estimate for the variance ϵ? This is a complicated question that is directly related to the study of the depth distortion function [4, 5, 11, 19]. However, the parameter γ necessary for defining the error in Equation 5.9 is the ratio of the tracking error, and the depth estimation error. This ratio is independent of the absolute magnitude ϵ.

5.2.3 Linking Algorithm

Suppose now that we have known corresponding points in two viewpoints. This allows the definition of an error measure that determines how well Equation 5.9 can be satisfied for these points and some equiform transformation. This error is computed for every pair of translations, one from each viewpoint along this valley. The best pair defines the camera translation for each camera. Figure 5.2, shows the error function for all pairs of translations chosen from along the valleys in each viewpoints.

Formally,

- For all pairs $(\hat{\mathbf{t}}_j^{(1)}, \hat{\mathbf{t}}_k^{(2)})$, with $\hat{\mathbf{t}}_j^{(1)} \in V^{(1)}$, and $\hat{\mathbf{t}}_k^{(2)} \in V^{(2)}$
  - Create the differential reconstructions using $\hat{\mathbf{t}}_j^{(1)}$ and $\hat{\mathbf{t}}_k^{(2)}$, and use this differential reconstruction to calculate the corresponding scene point sets $P^{(1)}, P^{(2)}$

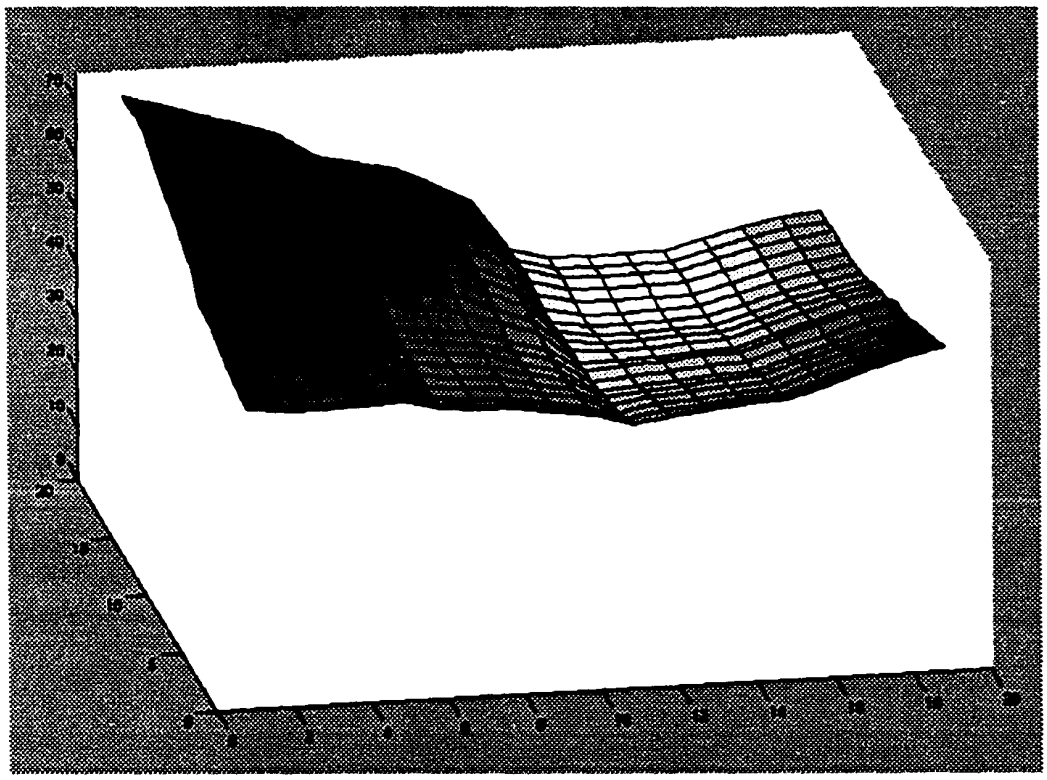


Figure 5.2: The x and y axes of the above figure are the positions along the ambiguous valley for each viewpoint. Each position along this axis defines a reconstruction made from different translations along the valley. The z axis represents the error: a measure of how well some equiform transformation can map points reconstructed from one viewpoint to (manually corresponded) points reconstructed from the other viewpoint.

- compute matching error:

$$g(j,k) = min_{\{R^{(1\to 2)},T^{(1\to 2)},s\}} \sum_i \left|\zeta(sR^{(1\to 2)}P_i^{(1)} + T^{(1\to 2)}) - \zeta(P_i^{(2)})\right|$$

- Let (j^*, k^*) be coordinates of the minimim error of g.
  - The "best translational estimates" are $\hat{\mathbf{t}}_{j^*}^{(1)}, \hat{\mathbf{t}}_{k^*}^{(2)}$
  - Use these translation estimates to compute $P^{(1*)}, P^{(2*)}$
  - The best equiform transformation relating the reconstructions is:

$$\{R^{(1\to 2)^*}, T^{(1\to 2)^*}, s^*\} =$$

$$argmin_{\{R^{(1\to 2)},T^{(1\to 2)},s\}} \sum_i \left|\zeta(sR^{(1\to 2)}P_i^{(1*)} + T^{(1\to 2)}) - \zeta(P_i^{(2*)})\right|$$

This process gives a method for choosing the correct translational velocity at each viewpoint. Figure 5.3 shows reconstructed depth maps for both viewpoints, showing the final depth map (on the left), and the depth map originally created with the translation vector that best fit the image data from only one viewpoint (on the right).

In order to make higher quality models of the scene structure, it is necessary to combine texture data from each viewpoint and correlate more than a sparse collection of matched points. The translation and rotation of the best equiform transformation: $(R^{(1\to 2)^*}, T^{(1\to 2)^*})$ is the rigid motion relating the two camera viewpoints. This allows the use of standard stereo algorithms to compute the depth maps. We use texture based stereo with normalized cross correlation using the differential reconstruction to give an initial depth map. Figure 5.3 (on the

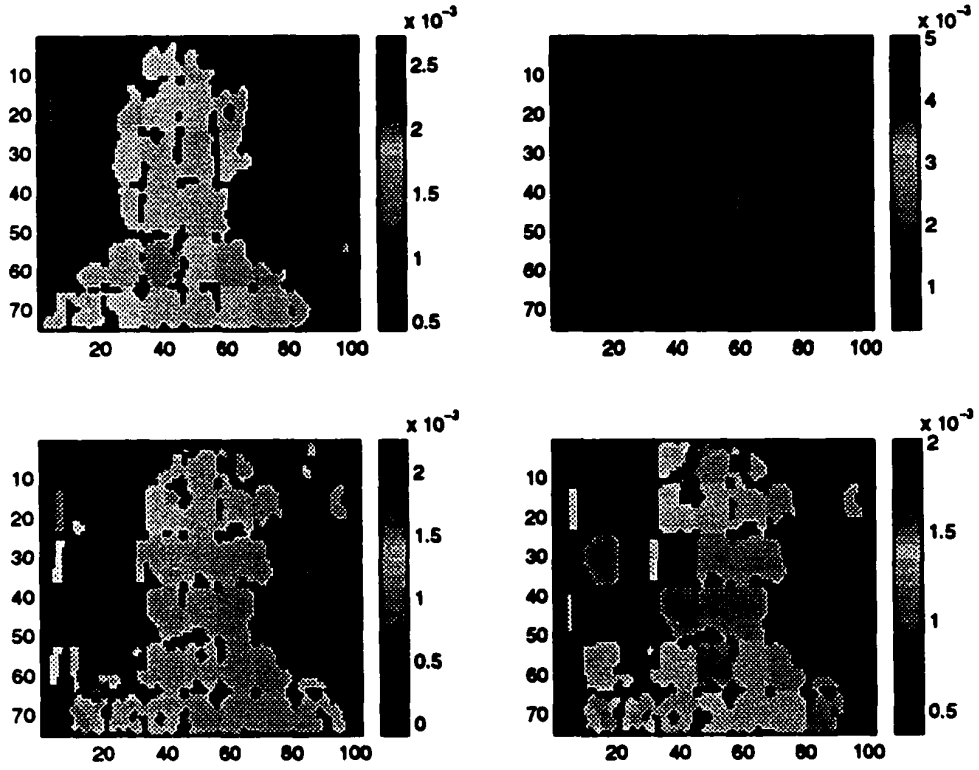


Figure 5.3: On the right are the depths maps for the lowest error translation in each frame. When all pairs of plausible translations for each frame are combined, the best pair gives the reconstructions shown on the left.

top left) shows the differential reconstruction starting condition. Patches with approximately the same depth were combined to define a mesh structure. The mesh vertices were then optimized based on normalized cross correlation of the textured model projected into the second image and the second image itself. Figure 5.4 shows the final result from new and quite different viewpoints.

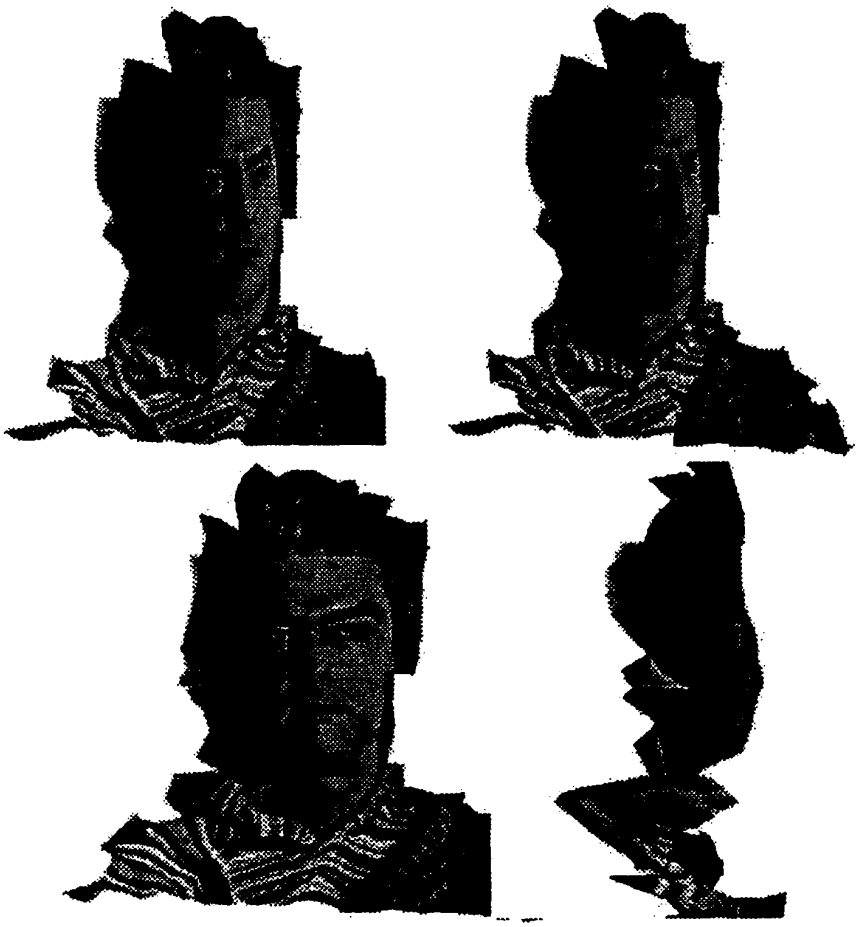


Figure 5.4: After finding the Equiform Transformation between viewpoints, standard stereo techniques can accurately fix the depth maps. In this case, a continuous mesh was defined over the head region automatically segmented using the techniques in Section 4.1. Texture based stereo using normalized cross correlation was use to optimize the mesh depths, using the differential reconstruction as an initialization.

Bibliography

[1] E. H. Adelson and J. A. Movshon. Phenomenal coherence of moving visual patterns. *Nature*, 300:523–525, December 1982.

[2] G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:477–489, 1989.

[3] Y. Aloimonos and Z. Duric. Estimating the heading direction using normal flow. *International Journal of Computer Vision*, 13:33–56, 1994.

[4] G. Baratoff. *Qualitative Space Representations Extracted from Stereo.* PhD thesis, Department of Computer Science, University of Maryland, 1997.

[5] G. Baratoff and Y. Aloimonos. Changes in surface convexity and topology caused by distortions of stereoscopic visual space. In *Proc. European Conference on Computer Vision*, pages II:226–240, 1998.

[6] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(2):167–192, 1988.

[7] T. Brodsky. *The Video Yardstick.* PhD thesis, Department of Computer Science, University of Maryland, 1999.

[8] T. Brodský, C. Fermüller, and Y. Aloimonos. Directions of motion fields are hardly ever ambiguous. *International Journal of Computer Vision*, 26:5–24, 1998.

[9] T. Brodský, C. Fermüller, and Y. Aloimonos. Self-calibration from image derivatives. In *Proc. International Conference on Computer Vision*, pages 83–89, 1998.

[10] D. Burke and P. Wenderoth. The effect of interactions betwen one-dimensional component gratings on two-dimensional motion perception. *Vision Research*, 33(3):343–350, 1993.

[11] L. Cheong. *The Geometry of the Interaction between 3D Shape and Motion.* PhD thesis, Department of Computer Science, University of Maryland, 1996.

[12] K. Daniilidis. *On the Error Sensitivity in the Recovery of Object Descriptions.* PhD thesis, Department of Informatics, University of Karlsruhe, Germany, 1992. In German.

[13] K. Daniilidis and M. E. Spetsakis. Understanding noise sensitivity in structure from motion. In *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles*, chapter 4. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[14] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *Proc. European Conference on Computer Vision*, pages 321–334, Santa Margherita Ligure, Italy, 1992.

[15] C. Fermüller. Passive navigation as a pattern recognition problem. *International Journal of Computer Vision*, 14:147–158, 1995.

[16] C. Fermüller and Y. Aloimonos. Direct perception of three-dimensional motion from patterns of visual motion. *Science*, 270:1973–1976, 1995.

[17] C. Fermüller and Y. Aloimonos. Qualitative egomotion. *International Journal of Computer Vision*, 15:7–29, 1995.

[18] C. Fermüller and Y. Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *International Journal of Computer Vision*, 28:137–154, 1998.

[19] C. Fermüller, L. Cheong, and Y. Aloimonos. Visual space distortion. *Biological Cybernetics*, 77:323–337, 1997.

[20] V. P. Ferrera and H. R. Wilson. Direction specific masking and the analysis of motion in two dimenions. *Vision Research*, 27:1783–1796, 1987.

[21] V. P. Ferrera and H. R. Wilson. Perceived speed of moving two-dimensional patterns. *Vision Research*, 31(5):877–893, 1991.

[22] K. J. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Proc. IEEE Image Understanding Workshop*, pages 156–161, 1991.

[23] R. Hartley. Euclidean reconstruction from uncalibrated views. In J. L. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, pages 237–256. Springer-Verlag, Berlin, 1994.

[24] D. Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1:279–302, 1988.

[25] J. Heel. Direct estimation of structure and motion from multiple frames. Technical Report MIT-AI-MEMO 1190, MIT, 1990.

[26] E. Hildreth. Computations underlying the measurement of visual motion. *Artificial Intelligence*, 23:309–354, 1984.

[27] T. J. Hine, M. Cook, and G. T. Rogers. An illusion of relative motion dependent upon spatial frequencies and orientation. *Vision Research*, 33(22):3093–3102, 1995.

[28] T. J. Hine, M. Cook, and G. T. Rogers. The Ouchi illusion: An anomaly in the perception of rigid motion for limited spatial frequencies and angles. *Perception and Psychophysics*, 59(3):448–455, 1997.

[29] B. K. P. Horn. *Robot Vision.* McGraw Hill, New York, 1986.

[30] B. K. P. Horn. Motion fields are hardly ever ambiguous. *International Journal of Computer Vision*, 1:259–274, 1987.

[31] B. K. P. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[32] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[33] B. K. P. Horn and E. J. Weldon, Jr. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.

[34] R. S. Jasinschi, A. Rosenfeld, and K. Sumi. Perceptual motion transparency: The role of geometrical information. *Journal of the Optical Society of America A*, 9:1865–1879, 1992.

[35] N. E. O. K. J. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Proc. International Conference on Computer Vision*, pages 357–365, 1993.

[36] D. H. Kelly. Motion and vision. II. stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America*, 609(10):1340–1349, October 1979.

[37] B.-G. Khang and E. A. Essock. Apparent relative motion from a checkerboard surround. *Perception*, 26(7):831–846, 1997.

[38] B.-G. Khang and E. A. Essock. A motion illusion from two-dimensional periodic patterns. *Perception*, 26(5):585–597, 1997.

[39] J. Kim and H. R. Wilson. Dependence of plaid motion coherence on component grating directions. *Vision Research*, 33(17):2479–2489, 1993.

[40] F. L. Kooi, K. K. D. Valois, D. H. Grosof, and R. L. D. Valois. Properties of the recombination of one-dimensional motion signals into a pattern motion signal. *Perception and Psychophysics*, 52(4):415–424, 1992.

[41] H. C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proc. Royal Society, London B*, 208:385–397, 1980.

[42] D. Marr and S. Ullman. Directional selectivity and its use in early visual processing. *Proc. Royal Society, London B*, 211:151–180, 1981.

[43] S. J. Maybank and O. D. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8:123–151, 1992.

[44] J. Mendelsohn, E. Simoncelli, and R. Bajcsy. Discrete-time rigidity constrained optical flow. In *Proc. 7th International Conference on Computer Analysis of Images and Patterns*. Springer, Berlin, 1997.

[45] H.-H. Nagel. Optical flow estimation and the interaction between measurement errors at adjacent pixel positions. *International Journal of Computer Vision*, 15:271–288, 1995.

[46] H.-H. Nagel and M. Haag. Bias-corrected optical flow estimation for road vehicle tracking. In *Proc. International Conference on Computer Vision*, pages 1006–1011, January 1998.

[47] K. Nakayama and G. H. Silverman. The aperture problem - I. Perception of nonrigidity and motion direction in translating sinusoidal lines. *Vision Research*, 28:739–746, 1988.

[48] K. Nakayama and G. H. Silverman. The aperture problem - II. Spatial integration of velocity information along contours. *Vision Research*, 28:747–753, 1988.

[49] S. Negahdaripour and B. Horn. Determining 3-d motion of planar objects from image brightness patterns. In *IJCAI85*, pages 898–901, 1985.

[50] J. Oliensis. A new structure from motion ambiguity. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–191, 1999.

[51] H. Ouchi. *Japanese and Geometrical Art*. Dover, 1977.

[52] C. Shu and Y. Shi. Direct recovering of nth order surface structure using unified optical flow field. *PR*, 26:1137–1148, 1993.

[53] D. Shulman and J.-Y. Hervé. Regularization of discontinuous flow fields. In *Proc. IEEE Workshop on Visual Motion*, pages 81–86, 1989.

[54] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 310–315, 1991.

[55] D. Sinclair, A. Blake, and D. Murray. Robust estimation of egomotion from normal flow. *International Journal of Computer Vision*, 13:57–69, 1994.

[56] A. T. Smith. Coherence of plaids comprising components of disparate spatial frequency. *Vision Research*, 32(2):393–397, 1992.

[57] A. T. Smith and G. K. Edgar. Perceived speed and direction of complex gratings and plaids. *Journal of the Optical Society of America*, 8(7):1161–1171, July 1991.

[58] L. Spillmann, U. Tulunay-Keesey, and J. Olson. Apparent floating motion in normal and stabilized vision. *Investigative Ophthalmology and Visual Science, Supplement*, 34:1031, 1993.

[59] S. Srinivasan and R. Chellappa. Noise-resilient estimation of optical flow by use of overlapped basis functions. *JOSAA*, 1999.

[60] G. R. Stoner, T. D. Albright, and V. S. Ramachandran. Transparency and coherence in human motion perception. *Nature*, 344:153–155, March 1990.

[61] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, 5(3):271–301, 1990.

[62] A. Verri and T. Poggio. Against quantitative optic flow. *Proc. IEEE International Conference on Computer Vision*, 1987.

[63] H. Wallach. Uber visuell wahrgenommene bewegungsrichtung. *Psychologische Forschung*, 20:325–380, 1935.

[64] S. Wang, Y. Markandey, and A. Reid. Total least squares fitting spatiotemporal derivatives to smooth optical flow fields. In *Proc. of the SPIE: Signal and Data Processing of Small Targets*, volume 1698, pages 42–55. SPIE, 1992.

[65] J. Weber and J. Malik. Robust computation of optical flow in a multi-scale differential framework. *International Journal of Computer Vision*, 14:67–81, 1995.

[66] L. Welch. The perception of moving plaids reveals two motion-processing stages. *Nature*, 337:735–737, Feb 1989.

[67] J. Weng, T. S. Huang, and N. Ahuja. *Motion and Structure from Image Sequences.* Springer-Verlag, Berlin, 1991.

[68] C. Yo and H. R. Wilson. Moving 2D patterns capture the perceived direction of both lower and higher spatial frequencies. *Vision Research*, 32:1263–1270, 1992.

[69] G. S. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:995–1013, 1992.