ERROR CHARACTERISTICS OF SFM WITH UNKNOWN FOCAL LENGTH

XIANG XU

(*B. Eng. Tianjin University*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2006

Acknowledgments

I would like to express my appreciation to Associate Prof. Cheong Loong Fah and Prof. Ko Chi Chung for their advice during my doctoral research endeavor for the past four years. As my supervisors, they have constantly forced me to remain focused on achieving my goal. Their observations and comments helped me to establish the overall direction of the research and to move forward with investigation in depth.

I also wish to thank my colleagues and friends at the National University of Singapore for always inspiring me and helping me in difficult times.

My family have given me a lot of love and support throughout the years. Their love, patience and sacrifice have made all of this possible.

Contents

Summary

The structure from motion (SFM) problem has been studied extensively by the computer vision community in the past two decades. SFM amounts to the problem of recovering the structure of 3-D scene and the 3-D relative motion between the scene and the observer from the projection of the 3-D relative motion onto a 2-D surface. If the camera is calibrated, camera motion can be recovered and Euclidean reconstruction of the scene can be carried out. While many algorithms have been developed for camera calibration, most are sensitive to noise and lack robustness and reliability.

In this thesis we present a theoretical analysis of the behavior of SFM algorithms with respect to the errors in intrinsic parameters of the camera. In particular, we are concerned with the limitation of SFM algorithms in the face of errors in the estimation of the focal length. This is important for camera systems with zoom capability and online calibration cannot be always done with the requisite accuracy. The results show that the effect of erroneous focal length on the motion estimation is not the same over different translation and rotation directions. The structure of the scene (depth) affects the shifting of the motion estimate as well. Simulation

with synthetic data and real images was conducted to support our findings.

We also attempt to explain the paradox of the unnoticed distortions when viewing the cinema. Cinema viewed from a location other than its Canonical Viewing Point (CVP) presents distortions to the viewer in both its static and dynamic aspects. Past works have investigated mainly the static aspect of the problem and attempted to explain why viewers still seem to perceive the scene very well. The dynamic aspect of depth perception has not been well investigated. We derive the dynamic depth cues perceived by the viewer and use the iso-distortion framework to understand its distortion. The result is that viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems. Despite this shift, the key properties of the perceived depths remain largely the same, being determined in the main by the accuracy to which extrinsic motion parameters can be recovered. And for a viewer seated at a non-central position and watching the movie screen with a slant angle, the view is related to the view at the CVP by a homography, resulting in various aberrations such as non-central projection.

List of Figures

Chapter 1

Introduction

1.1 What this thesis is about

The problem of inferring 3-D information of a scene from a set of 2-D images has a long history in computer vision. Although the basic geometric relationships governing the problem of structure and motion recovery from image sequences are well understood, the task is still unsolved and formidable. The reason for this half-failure is that, by its very nature, this problem falls into the category of so-called inverse problems, which are prone to be ill-conditioned and difficult to solve in their full generality unless additional assumptions are imposed. Despite these negative remarks, there has been a rapid development in computer vision over the two past decades. In particular, the Structure from Motion (SFM), which is defined as the

extraction of 3-D structure of a moving scene from image sequence, has become the central topic of computer vision community and received increasing attention. Since the existing SFM algorithms are very sensitive to noise, there have been many error analyses in the literature. In this thesis, we propose an approach to understand the detailed nature of the inherent ambiguities caused by the geometry of the problem itself and thus cannot be removed by any statistical schemes.

The problem of SFM is usually divided into three steps: (1) extract features and match them between images, (2) estimate the 3-D relative motion (ego-motion or object motion) and (3) recover depth or structure based on the results of the first two steps. Since both the recovery of 3-D motion from image motion, and the image motion estimation process are ill-posed in nature, SFM is difficult to solve robustly. Thus to understand the error characteristics of SFM algorithms is critical not only for knowing the limitations of the existing algorithms, but also for developing better algorithms. We take a step towards this direction. Our results show that the effect of erroneous focal length on the motion estimation is not the same over different translation and rotation directions. The structure of the scene (depth) affects the shifting of the motion estimate as well.

The results are used to understand one paradox that has received extended interests from psychophysics researchers—the unnoticed distortions under cinematic viewing condition. That is, picture or cinema viewed from a location other its composition point or center of projection (CoP) should present distortions to the viewer in both

the static and dynamic aspects. However, picture or cinema viewing is apparently not limited to the location at the CoP. Many other positions can serve as reasonable viewpoints allowing layout to appear relatively normal. Many psychophysics and vision researchers proposed their approaches to this paradox. However, most of the hypotheses mainly attempt to deal with the static aspect of the problem. Our work focuses on the dynamic aspect of cinematic perception and investigates its distortion to be expected theoretically, by adapting the computational model of the SFM process.

The remainders of this chapter overview the motivating factors, study scope and contributions of our research. We close this chapter with the organization of the thesis.

1.2 Background

1.2.1 Overview of SFM

The longstanding efforts of human to understand the image formation process can be found in ancient civilizations throughout the world. However, the first work that is directly related to multiple-view geometry is attributed to Kruppa [53]. He proved that two views of five points are sufficient to determine both the relative

transformation between the views and the 3-D location of points up to finitely many solutions. The origin of a modern treatment is traditionally attributed to Longuet-Higgins [60], who in 1981 first proposed a linear algorithm for structure and motion recovery from two images of a set of points, based on the so-called epipolar constraint. This work proved the existence of the solutions for 3-D scene reconstruction from 2-D displacement and triggered many researchers to develop practical computer vision algorithms. Tsai and Huang [103] proved that given an essential matrix associated with the epipolar constraint, there are only two possible 3-D displacements. The study of the essential matrix then led to a three-step SVD-based algorithm for recovering the 3-D displacement from image correspondences.

The essential matrix approach based on the epipolar constraint recovers only the discrete 3-D displacement. Mathematically, the epipolar constraint works well only when the displacement between the two images is relatively large, i.e. large base-line are required. However, in real-time applications, even if the velocity of the moving camera is not small, the relative displacement between two consecutive images might become small due to the high frame rate. In turn, the algorithms become singular due to the small translation and the estimation results become less reliable. Thus, a differential version of the 3-D motion estimation problem is to recover the 3-D velocity of the camera from optical flow, developed from which the structure (depth) of the scene can be estimated. Although some algorithms address the problem of motion and structure recovery simultaneously [99], most techniques

try to decouple the two problems by estimating the motion first, followed by the structure estimation. In this thesis, we also view the two as separate problems.

Due to the inverse nature of the problem, the estimation of 3-D motion based on 2-D displacement is noise sensitive. A small amount of error in image measurements can lead to very different solutions. SFM algorithms proposed in the past two decades faced this problem to varying extent. Many error analyses [1, 24, 111] has been reported. Most of these analyses deal with specific algorithms each using different optimization techniques. In [75], Oliensis argues that theoretical analyses of algorithm behavior are crucial. These analyses should underlie any particular algorithms. It is important not only for understanding algorithms' properties, but also for conducting good experiments and for developing the best algorithms. In this thesis, we propose an approach that lends itself towards understanding the behaviors of SFM algorithms under a wide range of motion-scene configurations. We study one class of algorithms based on the weighted differential epipolar constraint which is adopted by most of the existing differential SFM algorithms using optical flow as input. The optimization proposed by Xiang and Cheong [110] is adopted in our work, since it permits an unifying view of these different algorithms. It is based on the difference between the original optical flow and the reprojected flow obtained via a backprojection of the reconstructed depth, analogous to the distance between the observation and reprojection of the recovered structure in the discrete case [113, 112].

If the intrinsic parameters of the camera are unknown, the SFM problem can only be "solved" under an uncalibrated scenario from which only projective structure can be recovered. Most studies [29, 40, 68, 81] conducted have dealt with the discrete case. If one wants to obtain the Euclidean structure, camera calibration must be carried out. Camera calibration in this thesis refers to the process of estimating the intrinsic parameters of the camera.

Similar to the SFM algorithms, calibration algorithms are also sensitive to noise. The process of camera calibration introduces additional errors in the measurements, which affect the final estimates of the motion and structure. This is the case both when the camera is calibrated off-line or when self-calibration techniques are used. With the exception of few, the study of these effects has not received much attention. In the discrete setting, Bougnoux [5] analysed the stability of the estimation of intrinsic parameters and their effects on structure estimation. In [38], Grossmann derived the covariances of the parameters of an uncalibrated stereo system with fixed calibration parameters and under the hypothesis that an a priori quality of the final estimates was showed in the context of nonlinear optimization techniques. The effects of calibration errors on the motion estimates in the discrete setting are explored by Svodoba and Sturm [94]. They derived the relations between noise in the camera parameters and the acceptability of the translation vector. They also found that the estimation of the rotation is very sensitive to the accuracy in the calibration parameters. We derived similar result using a geometrical perspective.

We also find that the effect of erroneous intrinsic parameters estimates on the motion estimation is not the same over different translation and rotation directions. Furthermore the structure of the scene and the field of view (FOV) of the camera affect the motion estimates as well.

1.2.2 The paradox of unnoticed distortion in slanted images

The puzzle of unnoticed distortions in slanted images was first addressed by La Gournerie in 1859 [79]. The paradox occurs in two forms. The first concerns viewing pictures either nearer of farther than the CoP but along the line extended between that point and (usually) the center of the picture; the second, and by far the more interesting and complex, concerns viewing pictures from the side at any distance. Both of these forms can happen in the cinema viewing scenario.

Several explanations have been offered for the apparent invariance of perceived layout and shape in pictures with changes in viewing position. One (perhaps dominant) view is that observers somehow actively (though perhaps unconsciously) "correct" or "compensate" for the perspective distortions of the retinal image due to oblique viewing. This typically involves a simultaneous awareness of the pictorial cues and the cues that reveal the structure of the picture surface. Cutting [21] argues that the slant at which pictures are viewed is usually small, and consequently

the distortions of the retinal image are too small to be noticed. Perkins [77] claims that such invariance is a byproduct of the viewer's expectations with known shapes. For example, if the retinal image is similar to the image that would be created by a cube, prior expectations force the percept to that of a cube. The invariance thus comes from the viewer's experience with object whose shapes are familiar or usually follow certain rules (right angles, parallel sides, symmetry). A third explanation claims that the invariance is the consequence of altering or re-interpreting the retinal image by recovering the position of the screen surface. For example, it is known [8] that the locations of three mutually orthogonal vanishing points in the visual field are sufficient to recover the CoP. Banks et al. [3] argues that a local slant mechanism is used to estimate the foreshortening due to viewing obliqueness and then adjust the percept derived from the retinal image to undo the foreshortening. For a more detailed review please refer to Chapter 4.

Unlike the previous approaches, we are concerned with the dynamic cues in cinema in our work. This is important because distortions are present in both its static and dynamic aspects. As testified by the original names of kinetoscope and moving pictures, cinema was understood from its birth as the art of motion. Motion dynamically changes the viewing perspectives of the spectators. Therefore, motion cues should be a privieged object of investigation. Our research on the dynamic cues argues that viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems. Despite this shift, the key

properties of the perceived depths remain largely the same, being determined in the main by the accuracy to which extrinsic motion parameters can be recovered. For a viewer seated at a non-central position and watching the movie screen with a slant angle, the view is related to the view at the CVP by a homography, resulting in various aberrations such as non-central projection.

1.3 Contributions

We summarize the major contributions of this thesis as below:

3-D motion estimation with erroneous intrinsic parameters We use the unified optimization criteria based on the differential epipolar constraint to analysis the effect of calibration errors on motion estimation. We show the effects of erroneous intrinsic parameters on motion estimation are determined not only by the errors in the intrinsic estimates, but are also related to the extrinsic parameters, i.e., the direction of the translational and rotational velocity.

Cinema viewing paradox We prove that the cinema viewed from a location other than the CoP is no more complex than an uncalibrated SFM problem, where in particular the focal length is fixed but potentially unknown. The only difference with the usual SFM problem is that the principal point offset can be very much larger than one usually encounters in such problem. The

changes caused by the large principal point offset in the characteristics of the depth distortion are highlighted.

1.4 Organization

The remainder of this thesis is organized in four chapters, followed by appendices and a bibliography. The next chapter, Chapter 2, provides the background for the specific problems addressed in the thesis. We review the basic algorithms of SFM and highlight the relative merits of our work. The various optimization criteria used in SFM are also reviewed for both the discrete and differential case. To facilitate the discussion of depth perception we also revisit the iso-distortion framework which is first introduced in [10]. Notations and models utilized in this thesis are also introduced.

Chapter 3 presents a theoretical analysis of the behavior of SFM algorithms with respect to the errors in intrinsic parameters of the camera. How uncertainty in the calibration parameters gets propagated to the motion estimates is demonstrated both analytically and in simulation. Analyses of the behavior of SFM under various motion and scene configurations have been conducted.

In Chapter 4, we focus on the explanation of the unnoticed distortion of cinema viewed from a location other than the CoP. We first prove that the image formation

process can be treated as a SFM problem with a twist. That is, the changes caused by the location shift from the CoP can be analogized to a traditional uncalibrated SFM problem, only with minor modification. Then we show that the distortions caused by the shifting of position and the pose of the viewer do not alter the abilities of structure perception compared to the calibrated case, which in turn explains the paradox. Unlike the previous research, our approach is concerned with the dynamic aspect of the problem.

In the last chapter, we conclude our work and discuss future research directions. In particular, we discuss extending our research to the camera calibration problem. The appendices include a possible solution of the decomposition of the homography matrix introduced in Chapter 4.

Chapter 2

Models and Literature Review

Structure from motion (SFM) has been a very active area of computer vision in the past 20 years. The idea is to recover the shape of objects or scenes from a sequence of images acquired by a camera undergoing an unknown motion. Usually it is assumed that the scene is made up of rigid objects possibly undergoing some kind of Euclidean motion. The vision community extensively developed computer systems to exploit stereopsis or motion parallax. Most of such approaches can be classified as feature-based (discrete approach) or optical flow-based (differential approach) based. Other classification criteria include the number of input image (two views or multiple views), the implementation techniques (linear or nonlinear) and the underlying geometric constraint (epipolar constraint or depth-is-positive constraint). We briefly review the feature-based and flow-based approaches to

facilitate our further discussion.

2.1 Feature based SFM

In general, in a discrete approach, if the relative position and orientation of the two cameras are known, the 3D position of the imaged point can be easily computed by triangulation. The use of the epipolar geometry for the estimation of the relative orientation or motion was first proposed by Longet-Higgins [60] in the early eighties of the last century. The so-called essential matrix linearly constraints the feature points in the two images of the stereo pair:

$$x_1^T \mathbf{E} x_2 = 0, \tag{2.1}$$

where x_1 and x_2 are two corresponding feature points on two images, and E is the essential matrix.

The 8 points algorithm developed by the author has the appealing property of being linear. Relative rotation and translation of the cameras can be estimated by a factorization of the essential matrix. When the camera calibration is unknown the matrix derived by the constraint in equation (2.1) is called the fundamental matrix F. This can still be used to estimate motion and then structure but only up to a projective transformation [27, 10]. Despite its simplicity the 8 points algorithm has often been criticized for its excessive sensitivity to noise and lots of other techniques

have been developed. These are mostly based on the minimization of functions of the epipolar distances and usually require iterative optimization techniques. Beardsley and Zisserman [4] proposed an interesting technique that uses the weighted 8 points algorithm iteratively. At each stage the estimated essential matrix is used to calculate weights for the features used in the computation. Such weights are estimated by calculating the epipolar distances and then used in the next iteration. Excellent reviews of other weighted schemes can be found in [64, 112]. Hartley [41] showed that the performance of the 8 points algorithm can be drastically improved by renormalizing point feature coordinates. In his experiments he proved that the final performance is very similar to that of more advanced and complex algorithms.

2.2 Flow based SFM

In the differential setting, feature point correspondence in the discrete approach is replaced by optical flow. This is the velocity field of the image features

The estimation of optical flow is based on the image brightness constancy equation which states that the apparent brightness $I(x;t)$ of moving objects remains constant over time. This implies that:

$$\frac{dI}{dt} = \nabla_x I u + \frac{\partial I}{\partial t} = 0 \tag{2.2}$$

The differential SFM problem has also been explored by many researchers: an

algorithm was proposed in 1984 by Zhuang et al. [115] with a simplified version given in 1988 [116]; and a first order algorithm was given by Waxman et al. [108] in 1987. Most algorithms start from the basic bilinear constraint relating optical flow to the linear and angular velocities and solve for rotation and translation separately using either numerical optimization techniques [7] or linear subspace methods [45, 44]. Kanatani [51] proposed a linear algorithm reformulating Zhuang's approach in terms of essential parameters and twisted flow. However, in these algorithms, the similarities between the discrete case and the differential case are not fully revealed and exploited.

Although the differential 3-D motion and depth estimation algorithms are chosen as our subjects of study, our approach and the results are still applicable to a wider range of SFM algorithms, including the discrete approach.

2.3 Camera calibration

One of the major problem faced in computer vision applications is the calibration of camera. Camera calibration in this thesis is defined as the process of estimating the intrinsic parameters of the camera. It is a prerequisite for the Euclidean reconstruction from motion, for without camera calibration, SFM has to be generalized using a projective approach.

A camera is usually calibrated with one or more images of an object of known size and shape. A flat plate with a regular patten marked on it [31, 102] is commonly used for this purpose. Calibration in this way has the limitation of not being able to calibrate the camera online while executing a visual task. It is important to note that changes in the intrinsic parameters may be deliberate. An example is the change in the focal length of the camera in performing a zoom operation. Hence, in several applications, online calibration is desired and of practical interest.

Intensive study on the self-calibration has been conducted [81, 100, 68, 30]. The general principle behind most self calibration methods is based on the recovery of the absolute conic, which is invariant under rotations and translations, and independent of the camera pose. In the pioneering work of Maybank and Faugeras [68], the authors considered constraints on the intrinsic parameters, which arise from the rigidity of the camera motion and which are based on the epipolar geometry of two views. These constraints are known as Kruppa's equations. Nevertheless, methods based on these equations are plagued by inaccuracy due to high sensitivity to noise, and also suffer from convergence problem. In particular, critical motion sequences (CMS) [88, 89] will lead to multiple solutions in camera calibration. CMS has been systematically classified by Sterm [88] in the case of constant intrinsic parameters. This classification has been extended to more general calibration constraints, such as varying focal length [89].

Our work is concerned with the behavior of motion and structure recovery with

erroneous calibration of the intrinsic camera parameters. We show that the uncertainty in the focal length estimation propagates to the motion estimation in a complex manner. This propagation is influenced by the extrinsic parameters. The coupling of intrinsic and extrinsic parameters is algorithm-independent, as long as certain constraints (e.g. epipolar constraint) are involved in the algorithms.

2.4 Models

In this section, the notion of a perspective camera and the parameters associated with the model are introduced.

The pinhole model is the most commonly used model to solve camera-related problem. In this simple model, the camera performs a perspective projection of a point $\mathbf{P}$ in the 3-D world onto a pixel point $\mathbf{p}$ in the 2-D image plane through an optical center O, guided by the principles of geometrical optics. Figure 2.1 introduces the notation associated with the general projection process. The reference frame is attached to the optical centre at O. A world point $\mathbf{P} = (X, Y, Z)^T$ is projected to

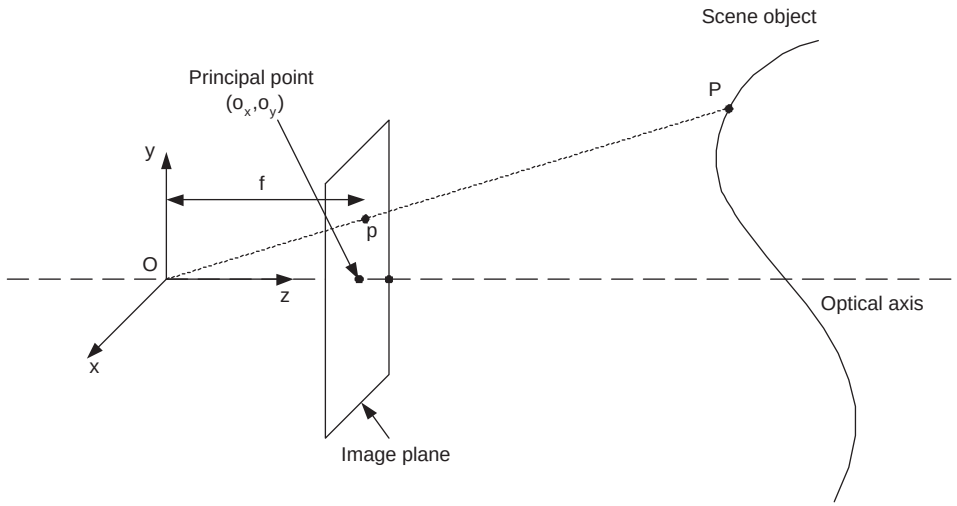


Figure 2.1: Image formation model: O is the optical centre. The optical axis is aligned with the Z-axis and the horizontal and vertical image axes are aligned with the X- and Y-axes respectively.

its image pixel coordinate (x, y) by the following well-known transformation [28]:

$$\mathbf{p} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \mathbf{K}\Pi_0\mathbf{P} = \begin{pmatrix} f & s_\theta & o_x \\ 0 & f & o_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2.3}$$

where we have expressed $\mathbf{p}$ and $\mathbf{P}$ in homogeneous coordinates, with slight abuse of notation in using $\mathbf{p}$ and $\mathbf{P}$ for both homogeneous coordinates and Euclidean coordinates. The constant 3×4 matrix Π_0 represents the perspective projection, and the upper triangular 3×3 matrix $\mathbf{K}$ is the intrinsic parameter matrix with the focal length denoted by f, (o_x, o_y) the x- and $y-$ coordinates of the principal point respectively, and s_θ the skew factor.

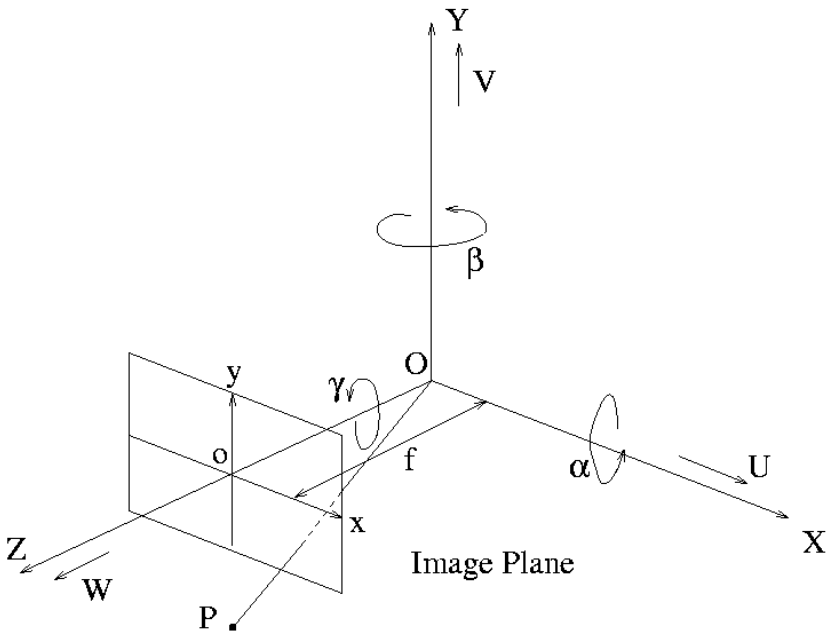


Figure 2.2: 3-D camera motion

We now present the notation associated with the conventional SFM problem, ignoring all the intrinsic parameters except f. It is equivalent to saying one has perfect estimates of the intrinsic parameters so that one can appropriately transform the image coordinates to obtain $s_\theta = o_x = o_y = 0$. If the camera undergoes a motion with a translational velocity $\mathbf{v} = (U,\ V,\ W)^T$ and a rotational velocity $\mathbf{w} = (\alpha,\ \beta,\ \gamma)^T$ (see Figure 2.2), the motion induces a relative motion between the static scene point $\mathbf{P}$ and the camera. The relative 3-D velocity of $\mathbf{P}$ (with respect to the camera) can be written as follows:

$$\dot{\mathbf{P}} = -\mathbf{v} - \mathbf{w} \times \mathbf{p}, \tag{2.4}$$

from which the well known 2-D motion field equations [60] can be derived:

$$u = \frac{W}{Z}x - f\frac{U}{Z} + \frac{xy}{f}\alpha - f\left(1 + \frac{x^2}{f^2}\right)\beta + \gamma y \tag{2.5}$$

$$v = \frac{W}{Z}y - f\frac{V}{Z} - \frac{xy}{f}\beta + f\left(1 + \frac{y^2}{f^2}\right)\alpha - \gamma x \tag{2.6}$$

where (u, v) is the optical flow at the feature point (x, y) on the image plane. We define $\dot{p}_{tr} = (u_{tr}, v_{tr})^T$and $\dot{p}_{rot} = (u_{rot}, v_{rot})^T$, where $\frac{\dot{p}_{tr}}{Z}$ and $\dot{p}_{rot}$ are the flows components due to translation and rotation respectively. Since only the translational direction can be recovered from the flow field, we can set $W = 1$ without loss of generality.

We introduce further notations for our distortion analysis. The estimated parameters are denoted with the hat symbol ($\hat{}$) and errors in the estimated parameters with the subscript e. The error of any estimate r is defined as $r_e = r - \hat{r}$.

2.5 Iso-distortion framework

The iso-distortion framework was first introduced by Cheong et al. [10]. The iso-distortion framework seeks to understand the geometric laws under which the recovered scene is distorted due to some errors in the estimated camera parameters. The distortion in the perceived space is visualized by looking at the locus of equal distortion, known as the iso-distortion surfaces. This makes explicit the systematic way in which depths are distorted and leads to its algebraic characterization by

Cremona transformation [48].

Referring to equations (2.5) and (2.6), we note that if there are errors in the estimates of the extrinsic parameters, these errors will in turn cause errors in the estimation of the scaled depth. The distorted depth $\hat{Z}$ can be shown to be given by:

$$\hat{Z} = Z\left(\frac{(x-\hat{x}_0, y-\hat{y}_0)\cdot\mathbf{n}}{(x-x_0, y-y_0)\cdot\mathbf{n} + Z\,(u_{rot_e}, v_{rot_e})\cdot\mathbf{n}}\right), \tag{2.7}$$

equation (2.7) shows that errors in the motion estimates distort the recovered relative depth by a factor D, given by the terms in the bracket, which among other terms, contains the term $\mathbf{n}$. The value of $\mathbf{n}$ depends on the scheme we use to recover depth. In our work, we choose to recover depth along the estimated epipolar direction, i.e. $\mathbf{n} = \frac{(x-\hat{x}_0, y-\hat{y}_0)^T}{\sqrt{(x-\hat{x}_0)^2+(y-\hat{y}_0)^2}}$. Such a choice is reasonable because the estimated epipolar direction contains the strongest translational flow and hence is the most reliable direction to recover Z. Hence the distortion factor D becomes:

$$D = \frac{(x-\hat{x}_0)^2 + (y-\hat{y}_0)^2}{(x-x_0, y-y_0)\cdot(x-\hat{x}_0, y-\hat{y}_0) + Z\,(u_{rot_e}, v_{rot_e})\cdot(x-\hat{x}_0, y-\hat{y}_0)}. \tag{2.8}$$

The complexity of equation (2.8) can be better grappled with a graphical approach in its first analysis. For specific values of the parameters $x_0,\ y_0,\ \hat{x}_0,\ \hat{y}_0,\ \alpha_e,\ \beta_e,\ \gamma_e$ and for any fixed distortion factor D, equation (2.8) describes a surface $g(x, y, Z) = 0$ in the xyZ-space. Normally, under general motion, a complicated distortion characteristic may arise. Readers are referred to [11, 12] for a full description of the geometry of the distortion.

Algebraically, it was shown from [10] that the transformation from physical to perceptual space belongs to the family of Cremona transformations. Such transformation is bijective almost everywhere except on the set of what is known as fundamental elements where the correspondence between the two spaces becomes one-to-many [48]. The complex nature of this transformation makes it clear that in general it is very difficult to recover metric depth accurately. What is less clear is the feasibility of recovering some of the less metrical depth representations under specific motions. For instance, the ordinal representation of depth constitutes one such reduced representation of depth where only depth order is available. Cheong and Xiang [12] showed that though in the general case, small amount of motion errors can have significant impact on depth recovery, there exist generic motions that allow robust recovery of partial depth information. In particular, lateral motion is better than forward motion in terms of yielding ordinal depth information and other aspects of depth recovery. On the other hand, forward motion leads to condition more conducive for 3-D motion estimation than that presented by lateral motion.

In the case of uncalibrated motion with fixed intrinsic parameters and reasonably small principal point offset, the distortion factor D becomes [12]:

- for lateral motion:

$$D = \frac{\hat{f}\hat{U}}{fU + \left(\beta f - \hat{\beta}\hat{f}\right) Z}, \tag{2.9}$$

- for forward motion:

$$D = \frac{x^2 + y^2}{((x - o_{xe})\, x + (y - o_{ye})\, y) + \left(-\left(\beta f - \hat{\beta}\hat{f}\right) x + \left(\alpha f - \hat{\alpha}\hat{f}\right) y\right) Z}. \tag{2.10}$$

It was shown in [12] that the aforementioned properties regarding depth and motion recovery are not affected, in spite of possible errors in the intrinsic parameters. However, if the intrinsic parameters are allowed to vary dynamically, then ordinality of depth will be lost under lateral motion.

The upshot of characterizing depth distortion behaviour under these generic types of forward and lateral motions are the following two aspects. (1) It shows that the reliability of a reconstructed scene has quite a different behaviour from that of the motion estimates. For instance, if the motion contains dominant lateral translation, it might be very difficult to lift the ambiguity between translation and rotation. However, in spite of such motion ambiguity, certain aspect of depth information seems recoverable with robustness. Indeed, in the biological world, lateral motions are often executed to judge distance and relative ordering. On the other hand, psychophysical experiments [104] reported that under pure forward translation, human subjects were unable to recover structure unless favorable conditions such as large field of view exist. Thus it seems that not all motions are equal in terms of robust depth recovery and that there also exists certain dichotomy between forward and lateral translation as far as motion and depth recovery are concerned.

(2) Understanding the depth recovered under these two very different motion types allows us to better able understand the behaviour of depth reconstruction under general motions, in the sense that the behaviour of depth reconstruction at the two opposite poles of translational motion spectrum delimits the type of general depth distortion behaviour somewhere in between the two poles.

2.6 SFM with erroneous estimation of intrinsic parameters: a literature review

Analysis of the theoretical precision of SFM estimates is common in photogrammetry, and increasing interaction between the computer vision and photogrammetry has resulted in an excellent synthesis [101] of photogrammetric bundle adjustment techniques which estimate jointly optimal 3-D structure and viewing parameter estimates. The survey highlights issues which might result in ill-conditioning and erratic numerical behaviour, such as a local parameterization that is nonlinear or excessive correlations, and unrealistic noise distributional assumption. Of course, much SFM error analysis has been done in the computer vision community [1, 24, 111]. Various ambiguities such as bas-relief ambiguity and opposite minimum were reported in the literature and were mainly attributed to the presence of noise in the image measurements [1, 24, 15]. Although dealing with the statistical

adequacy of the optimization criteria is important for understanding the effect of noise, it is equally important to understand the detailed nature of the inherent ambiguities caused by the geometry of the problem itself and thus cannot be removed by any statistical schemes. In [110], Xiang and Cheong argued that all the major ambiguities are actually inherent to the optimization criteria adopted and thus are algorithm-independent and will persist even with noiseless input. Oliensis [74] noted that in two-frame SFM, depth reconstruction from lateral motion suffers from the bas-relief ambiguity under which it is difficult to recover the constant component of the inverse-depths. They also found that under the more difficult situation of small range of depths and small translational baselines, the two-frame algorithm was more likely to encounter local minima when the true motion was forward than when it was sideways. Ma et al. [66] also examined the opposite minimum but termed it as the second eigenmotion. They noted that the opposite minimum can be distinguished from the true solution by using the positive depth constraint. Similar observations were made by [15, 32, 110] .

In recent years, there have been developments that result in continuing interest in SFM error analysis. One such development is the increasing variety of new camera models being proposed and considered [71, 91]. Pless [80] used the framework of the Fisher Information Matrix to understand how such standard rotation-translation ambiguity is modified in the case of multiple cameras arranged in different configurations. On the other hand, the widespread availability of video material recorded

with a zoom lens has also prompted investigation into different uncalibrated SFM algorithms and their related robustness properties. Errors in the intrinsic parameters might affect 3-D motion and scene recovery and various video applications such as 3-D virtual content insertion might be severely affected. Our work aims at unraveling the changes to the rotation-translation ambiguity that take place when there are uncertainties in the intrinsic parameters; it also discusses other associated properties such as the opposite minimum and the ordinality of recovered depths.

In the computer vision community and increasingly so for the photogrammetry community, there is a need to deal with uncalibrated camera. The computer vision community has developed schemes for self-calibration and investigated what structures can be recovered. Often in some applications such as object tracking, the real-time constraint of the application might mean that we are not able to calibrate the camera to the requisite degree of accuracy (e.g., ignoring the radial distortion) or that we have to ignore small changes or errors in the intrinsic parameters so that the computational complexity can be reduced. For instance, the quasi-Euclidean approach [17] computes the plane at infinity based on an approximate calibration of the intrinsic parameters. Indeed, in some applications a full-fledged Euclidean reconstruction is not necessary, for instance in visual servoing or in image-based rendering. Projective approaches aim to perform SFM without calibration, that is all the calibration information is neglected and the intrinsic camera parameters are assumed to vary freely from frame to frame. Oliensis [75] questioned whether the

projective approach might not be too general to a fault. The projective approach assumes zero knowledge of the calibration. In practice, there are always something we may say about the intrinsic camera parameters. Certain parameters might be known, such as the skew factor being zero, or we whether have a rough estimate of a certain parameter even though it might not be exact. It is questionable whether such neglect of available information leads to an increased or decreased robustness. However, despite the enormous amount of done work on developing projective algorithms, we still do not know when the projective approach is the right tool for its main task of dealing with calibration uncertainty. To answer this, we need to know how such simplification might affect the estimation of the camera's egomotion and, accordingly, scene recovery, and whether these influences are large enough in practice to affect the goal of tasks to be carried out by the camera system.

Bougnoux [5] noted the difficulty of obtaining focal length in self-calibration, but suggested from empirical evidence that part of the structure can be recovered despite error in self-calibration. The ground for this view, in so far as can be ascertained, seems to be based on the empirical results of depth reconstructed as part of the self-calibration process. Zhang [114] performed self calibration with a moving stereo rig, thus achieving redundancy compared to monocular sequence. They also empirically found that depths reconstructed are of good quality despite error in focal length estimate, but the depths are reconstructed using triangulation from the stereo pair (Type I measure in Tsai). Both did not address the accuracy of depth

reconstructions under general motion-scene configurations using those erroneous intrinsic parameters. Despite some works that suggest that depth reconstruction is stable against error in calibration, there is still a paucity of theoretical evidence that this notion is true for all motion-scene configurations.

Beside the work of Bougnoux [5], there have been many other works which report on the fact that self-calibration algorithms are sensitive to noise and lack robustness and reliability. Various researchers have analyzed the theoretical precision of the intrinsic estimates. Various authors [50, 88, 90, 59] analyzed the critical motion sequence in which no unique calibration can be obtained. Some of these general results have practical importance for certain motions and the special case of two-frame situation, which is also analyzed by Newsam et al. and Kahl and Triggs [50]. The significance of these works lies in that those configurations near to the critical motion sequence would yield unstable intrinsic estimates. However, how these uncertainties in estimating intrinsic parameters would in turn affect egomotion estimates is not made clear in these papers.

Other sources of errors arise from various simplifications and inaccuracies in the calibration process, and have indeed been the subject of various analysis. Lai [55] analyzed how the estimation of camera orientation and position would be affected when the offset of image centre and lens distortion are not included in the calibration process. Similar analysis on the role of lens distortion were carried out by [85, 109]. Lavest et al. [56] examined the influence of errors induced by the

metrology of calibration points on the accuracy of the intrinsic parameters. The results of these investigations support the notion that the effects of terms such as distortion and image offset seem to be minimal and can be left out for a simplified model.

Svoboda and Sturm [94] studied how uncertainty in the calibration parameters gets propagated to the motion parameters. Our work is closest in spirit to [94] in that it examines the effects of the intrinsic parameters on the estimation of the extrinsic parameters. However, instead of a statistical approach (as in [94]), we adopt a geometrical approach. The conclusions from [94] regarding the impact on the rotational component of the egomotion estimates are not clear, though it seems that the rotational estimates can be quite badly affected. The authors noted that the influence of the precision of the calibration parameters on the motion parameters estimation depends on the types of camera motion and the scene type. However, they did not further explore this scene-motion dependency. We investigate this dependency in a geometric manner and reveal further insights into this dependence.

Another work that investigated the coupling between the intrinsic and the extrinsic parameters is the recent work by González et al [36]. They have shown experimentally that there exists a strong coupling between the intrinsic and the extrinsic parameters. Most calibration methods, even those using static camera and calibration objects, suffer instability in the sense that that the set of intrinsic parameters returned by a calibration method suffered important variations under small dis-

placements of the camera relative to the calibration pattern. Similar results have been obtained for the extrinsic parameters when the camera only changed its internal configuration (i.e., when it zooms in or out) and not its relative position to the calibration pattern. In both cases the instability affects principally the parameters that are directly related with the element varied in the experiment. Therefore, when focal length varies the pattern distance is very unstable and vice versa. Similar results are obtained for the optic center when the pattern is displaced parallel to the camera. Although the error functions minimized by these different calibration techniques (usually minimizing the reprojection errors in the image or the reconstruction errors of the reference points in the 3-D space) yield similar error levels, it does not guarantee that the parameter estimates converge to the ground truth values, which is a serious problem if we want to use the calibrated camera in mobile applications. The main practical implication of this fact is that, when a camera is calibrated with any of these methods, we are "calibrating" the camera with just that pose. When subsequently the camera extrinsic parameters change, as in mobile applications, can we assume these "calibrated" intrinsic values for SFM analysis? Even the purported good quality of reconstructed depths under error in focal length estimate (as claimed by [5, 114]) might be true only with respect to that particular camera pose? In other words, one can get jointly optimal camera parameters (intrinsic and extrinsic) and depths in a calibration algorithm (optimal with respect to the cost function but without necessarily meaning that these camera parameters are correct), but when the camera pose changes, and without

calibrating again but rather using this fixed calibrated intrinsic parameters, the intrinsic parameters might be erroneous and these errors might affect the subsequent motion analysis. If errors in the intrinsic parameters indeed worsen the estimation of extrinsic motion parameters, depth reconstruction would be affected as the latter critically depends on accurate egomotion estimation, especially in certain scene-motion configuration such as forward translation [12]. Thus it is abundantly plausible that the task of depth reconstruction in the face of calibration uncertainty is a more complicated task than might be thought at first.

Finally, Oliensis [74] showed that unknown focal-length variations strengthen the effects of the bas-relief ambiguity. This is attributed to the simple fact that the zoom flow is essentially not recoverable from the forward translation component. Coupled with the rotation-translation coupling that gives rise to the original bas-relief valley, this new coupling renders all directions of the translation not accurately recoverable. The paper also went on to note that the motion errors depend simply on the estimated focal length and image center (e.g. the estimated translation differs from the true translations by factors of the unknown focal length), but this is based on various assumptions such as the non-translational terms can still be annihilated in the proposed algorithm and that second order terms are small. We look at this relationship between the extrinsic motion and the estimated focal length in detail, assuming that the camera is not undergoing zoom motion, and we found here that errors in the focal length modifies the phenomenon of bas-

relief ambiguity in a non-simple way and thus affects the determination of camera extrinsic motion.

Chapter 3

Error Characteristics of SFM with Unknown Focal Length

This chapter presents a theoretical analysis of the behavior of "Structure from Motion" (SFM) algorithms with respect to the errors in the intrinsic parameters of the camera. We demonstrate both analytically and in simulation how uncertainty in the calibration parameters gets propagated to motion estimates. We studied the behavior of the estimation of the focus of expansion (FOE) in the case that the camera is well calibrated except that the focal length is estimated with error. The results suggest that the behavior of the bas-relief ambiguity is affected by the erroneous focal length. The amount of influence depends on the relative direction of the translation and rotation parameters of the camera, the field of view and scene

depth. Simulation with synthetic data was conducted to support our findings.

3.1 Problem Statements

Much work about the SFM error analysis has been done in the last 15 years [1, 24, 111]. Various ambiguities such as bas-relief ambiguity and opposite minimum were reported in the literature and were mainly attributed to the presence of noise in the image measurements [1, 24, 15]. In [110], Xiang and Cheong argued that all the major ambiguities are actually inherent to the optimization criteria adopted and thus are algorithm-independent and will persist even with noiseless input. Although dealing with the statistical adequacy of the optimization criteria is important for understanding the effect of noise, it is equally important to understand the detailed nature of the inherent ambiguities caused by the geometry of the problem itself and thus cannot be removed by any statistical schemes. In this thesis, we adopt such geometrical approach and further the analysis of SFM with erroneous intrinsic calibration and uncalibrated scenario.

In a recent critique of SFM research, Oliensis [75] argues that more comprehensive theoretical as well as phenomenological analyses of algorithm behavior should be carried out under all sorts of typical scenarios. Such analyses are important not only for understanding algorithms' properties, but also for conducting good experiments

and for developing the best algorithms. Based on the work of [110], we propose in this thesis an approach that lends itself towards understanding the behavior of SFM algorithms in uncalibrated scenario. In particular, we are concerned with the limitation of SFM algorithms in the face of errors in the estimation of the focal length. This is important for camera systems with zoom capability, and online calibration cannot be always done with the requisite accuracy. Instead of dealing with specific algorithms, each using different optimization techniques, we study one class of algorithms based on the weighted differential epipolar constraint. It is based on the difference between the original optical flow and the reprojected flow obtained via a back projection of the reconstructed depth, analogous to the distance between the observed feature and the reprojection of the recovered structure in the discrete case. This criterion permits a unifying view of these different algorithms. It also allows us to develop a simple and explicit expression for the residual error in terms of the errors in the 3-D motion estimates and the intrinsic parameters and enables us to predict the exact conditions likely to cause ambiguities. The error surfaces under a wide range of motion-scene configurations are plotted, from which several results are drawn.

Like the SFM algorithms, calibration algorithms are also sensitive to noise and lack robustness and reliability. Given the difficult of calibrating the camera precisely, projective approaches aim to perform SFM without calibration, that is all the calibration information is neglected and the intrinsic camera parameters are

assumed to vary freely from frame to frame. Although in some applications a full-fledged Euclidean reconstruction is not necessary, for instance in visual servoing or in image-based rendering, the projective approach may be too general to a fault. Although enormous amount of work on developing projective algorithms have been carried out by researchers, we still do not know when the projective approach is the right tool for its main task of dealing with calibration uncertainty. The projective approach assumes zero knowledge of the calibration. In practice, there are always something we may say about the intrinsic camera parameters. It is questionable whether such neglect of available information leads to an increased or decreased robustness. To answer this, one thing we need to know is whether the calibration uncertainty is large enough in practice to affect the goal of motion estimation and depth reconstruction. Oliensis [75] reported that even small errors in the estimation of focal length led to significant errors in the 3-D motion estimation. In this thesis, we use the error surface to illustrate the behavior of egomotion estimation with erroneous calibration of the focal length.

If such an understanding can be achieved, we can better judge if there is a need of constant recalibration using robust but computationally intensive algorithms, or we can accept certain errors in the focal length estimate but at the same time are fully aware of the limit of the applicability of such algorithm. Due to space limitation, we assume in this thesis no errors in other intrinsic parameters. However the extension to those cases is not difficult and the results remain largely the same.

3.2 Optimization Criteria for SFM

Most of the existing cost functions for SFM are based on some forms of the epipolar constraint which was proposed by Longuet-Higgins [60]. The epipolar constraint relates the 3-D motion parameters with the image displacements in a manner independent of depth. In the discrete case, the SFM problem amounts to the estimation of the fundamental matrix $\mathbf{F}$ (or the essential matrix $\mathbf{E}$ in the calibrated case) based on a sufficiently large set of point correspondences [28] from the following epipolar equation:

$$\mathbf{p}_1\mathbf{F}\mathbf{p}_2 = 0 \tag{3.1}$$

where $\mathbf{p_1}$ and $\mathbf{p_2}$ are the corresponding image points in the two views. A couple of non-linear optimization criteria have been proposed to properly reflect the geometric meaning of the epipolar equation, namely, $\mathbf{p_1}$ must lie on the epipolar line of $\mathbf{p_2}$ given by $\mathbf{F}\mathbf{p_2}$ and $\mathbf{p_2}$ on the epipolar line of $\mathbf{p_1}$ given by $\mathbf{F^T}\mathbf{p_1}$. The most commonly used three criteria are respectively based on the distance between the observed point and its corresponding epipolar line (denoted by J_{D1}), the gradient-weighted epipolar error (denoted by J_{D2}) and the distance between the observed point and the reprojection of the reconstructed depth (denoted by J_{D3}). Zhang [112] studied the relationship between these three criteria under different motion configurations. J_{D2} was recommended since it is equivalent to the most optimal J_{D3} under most configurations and yet is computationally more efficient.

In the differential case, similar motion estimation algorithms can be developed based on the differential epipolar constraint. The epipolar equation in the differential case can be written as [6]

$$\mathbf{p}^T \bar{\hat{\mathbf{v}}} \dot{\mathbf{p}} + \mathbf{p}^{\mathbf{T}} \bar{\hat{\mathbf{v}}} \bar{\hat{\mathbf{w}}} \mathbf{p} = \mathbf{0} \tag{3.2}$$

from which one can minimize the following cost function

$$J_{E1} = \sum_{i=1}^{n} \left(\mathbf{p}_i^T \bar{\hat{\mathbf{v}}} \dot{\mathbf{p}}_\mathbf{i} + \mathbf{p}_\mathbf{i}^{\mathbf{T}} \bar{\hat{\mathbf{v}}} \bar{\hat{\mathbf{w}}} \mathbf{p}_\mathbf{i} \right)^2 \tag{3.3}$$

where n is the number of image velocity measurement. The constraint J_{E1} can also be written in the following equivalent form:

$$J_{E1} = \sum_{i=1}^{n} \left(\left([\dot{\mathbf{p}}_i]_2 - \hat{\dot{\mathbf{p}}}_{rot_i} \right) \cdot \hat{\dot{\mathbf{p}}}_{tr_i}^{\perp} \right)^2 \tag{3.4}$$

It says that in the image plane the derotated flow vector $[\dot{\mathbf{p}}_i]_2 - \hat{\dot{\mathbf{p}}}_{rot_i}$ should be parallel to the epipolar direction $\hat{\dot{\mathbf{p}}}_{tr_i}$, or equivalently perpendicular to $\hat{\dot{\mathbf{p}}}_{tr_i}^{\perp}$. However, a bias of the estimated translation is well-known to be present when a linear algorithm based on (3.4) is applied. In view of this bias, a statistically more adequate implementation of the differential epipolar constraint should be

$$J_{E2} = \sum_{i=1}^{n} \left(\frac{\left([\dot{\mathbf{p}}_i]_2 - \hat{\dot{\mathbf{p}}}_{rot_i} \right) \cdot \hat{\dot{\mathbf{p}}}_{tr_i}^{\perp}}{|| \, [\dot{\mathbf{p}}_i]_2 - \hat{\dot{\mathbf{p}}}_{rot_i} || \cdot || \hat{\dot{\mathbf{p}}}_{tr_i}^{\perp} ||} \right)^2 \tag{3.5}$$

Like the discrete case, there are a variety of other non-linear methods which are basically different weighted version of J_{E1}. In [110] a cost function which amounts to a weighted version of J_{E1} is proposed:

$$J_R = \sum_{i=1}^{n} \left(\frac{\hat{\dot{\mathbf{p}}}_{tr_i} \cdot \left([\dot{\mathbf{p}}_i]_2 - \hat{\dot{\mathbf{p}}}_{rot_i} \right)^{\perp}}{\hat{\dot{\mathbf{p}}}_{tr_i} \cdot \mathbf{n}_i} \right)^2 \tag{3.6}$$

where $\mathbf{n}_i$ is a unit vector in the image plane representing a particular direction associated with the i^{th} image point. Various weighted differential epipolar constraints differ mainly in the choice of this unit vector $\mathbf{n}$. It was also shown that the key properties of the various cost functions used in different algorithms are determined by the angle between the two vectors involved in the dot product in the numerator; the choice of $\mathbf{n}$ in the denominator might affect the detailed numerical properties but has little influence on key properties such as the formation of the bas-relief valley on the error surface.

3.3 Behavior of motion estimation algorithms with erroneous estimated focal length

The preceding section reviewed the general behavior of motion estimation algortihms for the calibrated case. This section will investigate the behavior of extrinsic motion estimation under erroneous camera calibration. In particular, we consider how extrinsic motion estimation would be affected by fixed errors in the estimates of the focal length and the principal point offset.

We have seen in the preceding section that how studying the error surface of J_R allows us to understand the behavior of SFM algorithms in an algorithm-independent way. Thus, we first need to express the cost function J_R in terms of the various

component errors in the 3-D motion estimates together with terms arising from errors in the estimates for the intrinsic parameters. For clarity of presentation, we first consider the case where the only intrinsic parameter with error is the focal length, leaving the full case to section 3.3.3. Substituting $\hat{\dot{\mathbf{p}}}_{\mathbf{tr_i}} = (x_i - \hat{x}_0, y_i - \hat{y}_0)^T$, $[\dot{\mathbf{p}}_\mathbf{i}]_2 = (u_i, v_i)^T = \left(\frac{x_i - x_0}{Z_i} + u_{rot_i}, \frac{y_i - \hat{y}_0}{Z_i} + v_{rot_i}\right)^T$ and $\dot{\mathbf{p}}_{\mathbf{rot_i}} = (\hat{u}_{rot_i}, \hat{v}_{rot_i})^T$ into Equation (3.6) we have:

$$J_R = \sum_{i=1}^{n} \left(\frac{(x_i - \hat{x}_0, y_i - \hat{y}_0) \cdot \left(v_{rot_e} - \frac{y_{0e}}{Z_i}, \frac{x_{0e}}{Z_i} - u_{rot_e} \right)}{(x_i - \hat{x}_0, y_i - \hat{y}_0) \cdot \mathbf{n_i}} \right)^2 \tag{3.7}$$

where the various error terms are expanded as follows:

$$\begin{aligned} (x_{0_e}, y_{0_e}) &= (x_0 - \hat{x}_0, y_0 - \hat{y}_0) \\ u_{rot_e} &= -\left(\beta f - \hat{\beta}\hat{f}\right) + \left(\frac{\alpha}{f} - \frac{\hat{\alpha}}{\hat{f}}\right) x_i y_i - \left(\frac{\beta}{f} - \frac{\hat{\beta}}{\hat{f}}\right) {x_i}^2 + \gamma_e y_i \\ v_{rot_e} &= \left(\alpha f - \hat{\alpha}\hat{f}\right) + \left(\frac{\alpha}{f} - \frac{\hat{\alpha}}{\hat{f}}\right) {y_i}^2 - \left(\frac{\beta}{f} - \frac{\hat{\beta}}{\hat{f}}\right) x_i y_i - \gamma_e x_i \end{aligned} \tag{3.8}$$

Besides the usual errors in the extrinsic motion parameters, new terms appear in the above expression due to the inaccurate focal length estimate $\hat{f}$. For notational convenience, we shall henceforth omit the subscript i in the expression of J_R, although it is understood that the summation runs over all feature points. Furthermore, we denote the terms in the numerator of Equation (3.7) $(x - \hat{x}_0, y - \hat{y}_0)^T$ and $\left(v_{rot_e} - \frac{y_{0_e}}{Z}, \frac{x_{0_e}}{Z} - u_{rot_e}\right)^T$ as $\mathbf{t}_1$ and $\mathbf{t}_2$ respectively, as in [110], and we will be analyzing how this angular relationship between $\mathbf{t}_1$ and $\mathbf{t}_2$ — the key to the formation of the bas-relief valley — will change in the light of calibration errors. We also

adopt the similar terminology that for the vectors $\mathbf{t}_1$ and $\mathbf{t}_2$, $\mathbf{t}_{1,n}$ and $\mathbf{t}_{2,n}$ denote the n^{th} order component with respect to x and y; thus we have:

$$\begin{aligned} J_R &= \sum \left(\frac{\mathbf{t}_1 \cdot \mathbf{t}_2}{\mathbf{t}_1 \cdot \mathbf{n}} \right)^2 \\ \mathbf{t}_1 &= \mathbf{t}_{1,0} + \mathbf{t}_{1,1} \\ \mathbf{t}_2 &= \mathbf{t}_{2,0} + \mathbf{t}_{2,1} + \mathbf{t}_{2,2} + \mathbf{t}_{2,Z} \end{aligned} \tag{3.9}$$

where

$$\begin{aligned} \mathbf{t}_{1,0} &= (-\hat{x}_0, -\hat{y}_0)^T \\ \mathbf{t}_{1,1} &= (x, y)^T \\ \mathbf{t}_{2,0} &= \left(\left(\alpha f - \hat{\alpha}\hat{f} \right), \left(\beta f - \hat{\beta}\hat{f} \right) \right)^T \\ \mathbf{t}_{2,1} &= (-\gamma_e x, -\gamma_e y)^T \\ \mathbf{t}_{2,2} &= \left(\left(\frac{\alpha}{f} - \frac{\hat{\alpha}}{\hat{f}} \right) y^2 - \left(\frac{\beta}{f} - \frac{\hat{\beta}}{\hat{f}} \right) xy, - \left(\frac{\alpha}{f} - \frac{\hat{\alpha}}{\hat{f}} \right) xy + \left(\frac{\beta}{f} - \frac{\hat{\beta}}{\hat{f}} \right) x^2 \right)^T \\ \mathbf{t}_{2,Z} &= \left(-\frac{y_{0e}}{Z}, \frac{x_{0e}}{Z} \right)^T \end{aligned}$$

Since the depth Z may be dependent on x and y in a complex manner, we use the notation $\mathbf{t}_{2,Z}$ without explicitly specifying the order of this term.

Equations (3.7) and (3.8) show that for any given data set (x, y, Z), the residual error is a function of the true FOE (x_0, y_0), the estimated FOE $(\hat{x}_0, \hat{y}_0)$, the error in the rotation estimates $(\alpha_e, \beta_e, \gamma_e)$ and the estimated focal length $\hat{f}$. In comparison with the calibrated case, we immediately note the following:

1. The estimation of γ is quite independent of camera calibration since the γ_e term is not coupled with the intrinsic parameters in any meaningful way. Thus, geometrically speaking, γ can be estimated well, like in the case of calibrated SFM.

2. Unlike the calibrated case where the cost function only depends on errors in the rotational parameters and not the true rotational parameters themselves (for the calibrated case, $\mathbf{t_{2,0}} = (\alpha_e f, \beta_e f)^T$ and $\mathbf{t_{2,2}} = (\alpha_e \frac{y^2}{f} - \frac{\beta_e xy}{f}, -\frac{\alpha_e xy}{f} + \beta_e \frac{x^2}{f})^T$), here the true rotational parameters do play a part in the formation of the error surface.

3.3.1 Changes to the Bas-Relief Valley

Clearly, as in the calibrated case [110], the properties of the motion estimation algorithms depend on the angular relationship between the terms in the numerator of equation (3.7). In particular, if there exists a class of motion solutions that make the dot product in the numerator vanish, then ambiguities exist. We recapitulate the two conditions discussed in [110] that should be satisfied to make the numerator of the cost function vanish:

$$\begin{aligned} &(1) \quad \text{making } \mathbf{t}_1 \text{ and } \mathbf{t}_2 \text{ perpendicular to each other, and} \\ &(2) \quad \text{making } ||\mathbf{t}_2|| \text{ small.} \end{aligned} \tag{3.10}$$

Condition (2) helps because condition (1) can never be completely satisfied at every image point under general motion-scene configuration with depth Z not a constant value. Making $||\mathbf{t}_1||$ small does not help since it appears in both the numerator and the denominator.

From the expressions of $\mathbf{t}_1$ and $\mathbf{t}_2$ in Equation (3.9), we can see that $\mathbf{t}_{1,0}$, $\mathbf{t}_{2,0}$ and $\mathbf{t}_{2,Z}$ are pointing towards constant directions for all the feature points. If we consider $\mathbf{t}_{1,1}$ as a perturbation to the constant-direction vector $\mathbf{t}_{1,0}$ and $(\mathbf{t}_{2,1} + \mathbf{t}_{2,2})$ as a perturbation to $(\mathbf{t}_{2,0}+\mathbf{t}_{2,Z})$,[1] then making the constant-direction vectors $(\mathbf{t}_{2,0}+\mathbf{t}_{2,Z})$ and $\mathbf{t}_{1,0}$ perpendicular to each other is a reasonable choice for the minimization of J_R. Thus we have

$$\frac{y_{0e} - \alpha f Z + \hat{\alpha}\hat{f}Z}{x_{0e} + \beta f Z - \hat{\beta}\hat{f}Z} = \frac{\hat{y}_0}{\hat{x}_0} \tag{3.11}$$

or equivalently

$$\frac{y_{0e} - \alpha_e f Z - \hat{\alpha} f_e Z}{x_{0e} + \beta_e f Z + \hat{\beta} f_e Z} = \frac{\hat{y}_0}{\hat{x}_0} \tag{3.12}$$

The last equation shows that, in the case $\hat{f} = f$, the last terms in both the numerator and the denominator on the left hand side vanish. The equation reduces to the calibrated case, and as discussed in [110] it can be satisfied by obeying two independent constraints, the first one relating to the translational parameters $\frac{x_{0e}}{y_{0e}} = \frac{\hat{x_0}}{\hat{y_0}}$

[1] This statement means that we require the feature points to be sufficiently evenly distributed such that the vectors $\mathbf{t_{1,1}}$ are evenly spread on either side of $\mathbf{t_{1,0}}$ and the sum of vectors $\mathbf{t_{2,1}}$ and $\mathbf{t_{2,2}}$ are evenly spread on either side of $\mathbf{t_{2,0}} + \mathbf{t_{2,Z}}$, and the distribution of depth Z is symmetrical with respect to the $\mathbf{t_{1,0}}$ direction.

(which implies $\frac{x_{0e}}{y_{0e}} = \frac{x_0}{y_0}$), and the second one relating to the rotational parameters $\frac{\alpha_e}{\beta_e} = -\frac{\hat{y_0}}{\hat{x_0}}$. The first constraint characterizes the valley that gives rise to the bas-relief ambiguity found in calibrated SFM algorithms. However, in the uncalibrated case, when the error in the focal length f_e is significant, α_e and β_e cannot be freely varied to satisfy the second constraint $\frac{\alpha_e}{\beta_e} = -\frac{\hat{y_0}}{\hat{x_0}}$. Rather, if there is significant error in the estimate $\hat{f}$, the term $\mathbf{t}_{2,2}$ can no longer be treated as second order effect and be ignored relative to $\mathbf{t}_{2,0}$. Comparing terms in $\mathbf{t}_{2,0}$ and $\mathbf{t}_{2,2}$, we observe that even if the FOV is small (i.e. f is large) such that $\frac{\alpha}{f} \ll \alpha f$ and $\frac{\beta}{f} \ll \beta f$, the corresponding relationships for the estimated terms $\frac{\hat{\alpha}}{\hat{f}} \ll \hat{\alpha}\hat{f}$ and $\frac{\hat{\beta}}{\hat{f}} \ll \hat{\beta}\hat{f}$ may not be valid, and thus we cannot assert $||\mathbf{t}_{2,2}|| \ll ||\mathbf{t}_{2,0}||$. This is the case when $\hat{f}$ is under-estimated such that the estimated FOV is large. Under such circumstance, making $||\mathbf{t}_{2,2}||$ small is just as important towards minimizing the cost function J_R. Clearly this gives rise to the following constraint on the rotational estimates:

$$\frac{\alpha}{f} = \frac{\hat{\alpha}}{\hat{f}}, \ \frac{\beta}{f} = \frac{\hat{\beta}}{\hat{f}} \tag{3.13}$$

Note that in the above, the quantity $\hat{f}$ is fixed (since we are considering fixed focal length estimate); thus the two equations in (3.13) fully specify $\hat{\alpha}$ and $\hat{\beta}$. With this constraint, α_e and β_e cannot be freely varied such that the original calibrated constraint $\frac{\alpha_e}{\beta_e} = -\frac{\hat{y_0}}{\hat{x_0}}$ is satisfied. Thus to satisfy constraint (3.11), we cannot decompose it into two independent constraints like in the calibrated case. Rather,

to satisfy both (3.11) and (3.13) at the same time, we substitute (3.13) into (3.11) and obtain a single constraint:

$$\frac{y_{0e} - \alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)}{x_{0e} + \beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)} = \frac{\hat{y}_0}{\hat{x}_0} \tag{3.14}$$

which can also be written as

$$\frac{y_0 - \alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)}{x_0 + \beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)} = \frac{\hat{y}_0}{\hat{x}_0} \tag{3.15}$$

The above expresses a constraint on the direction of the estimated FOE $(\hat{x}_0, \hat{y}_0)$ that dictates the formation of the bas-relief valley. Compared to the original bas-relief constraint in the calibrated case $\frac{\hat{y}_0}{\hat{x}_0} = \frac{y_0}{x_0}$, which is a straight line passing through the true FOE and the origin, this modified constraint indicates a "bas-relief" valley that has a different slope in general given by $\frac{y_0 - \alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)}{x_0 + \beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)}$. In particular, consider the shift in the FOE estimate $(\hat{x}_0, \hat{y}_0)$ caused by the term $\beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$ and $-\alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$. One can also interpret this shift as an additional bias to the FOE estimate caused by the error in the focal length estimate, over and above the well-known bias towards the optical center. This bias was also investigated in [65], but their approach has difficulty in analytically deriving the bias as a function of the various factors. Using simulation, they seemed to obtain the result that under-estimation of focal length results in a larger bias than over-estimation of focal length. We confirm and explain later that the bias is indeed larger for

under-estimation of focal length, but our approach also allows us to show how the direction of the FOE bias is a function of the actual translation and rotation.

Furthermore, recall from equation (3.10) that ambiguity is more likely to arise if $||\mathbf{t}_2||$ is also small. Of the terms in $||\mathbf{t}_2||$, the rotational errors α_e and β_e in $||\mathbf{t}_2||$ can no longer be freely varied due to equation (3.13); thus $\hat{x}_0$ and $\hat{y}_0$ are clearly constrained in magnitude in order to make $||\mathbf{t}_{2,Z}||$ and thus $||\mathbf{t}_2||$ small. In other words, $(\hat{x}_0, \hat{y}_0)$ is not only just constrained in direction but also in magnitude; this is unlike the small field calibrated case, where any residual error caused by the translational errors can be compensated for by a suitable choice of α_e and β_e. Accordingly, we expect in general that the bas-relief valley might not straddle across the entire visual field. In particular, the feasibility of the flipped minimum solution [75] that exists under calibrated scenario (i.e. $(\hat{x}_0, \hat{y}_0) = -(x_0, y_0)$) would be diminished. On the other hand, due to the presence of the Z term in the constraint (3.14), we expect the shape of this bas-relief valley to be markedly affected by the way the scene points are distributed. For a cluttered scene with non-smooth depth distribution, the valley will be less well-defined. That is, instead of a narrow and elongated valley that stretches across the entire visual field, it would be broader and rather reduced in length to a local quadrant. We also expect more local minima in the solution space due to the non-smooth Z term in the constraint (3.14), which could pose convergence problem for a Euclidean SFM algorithm assuming erroneous calibration parameters. As a result, using a projective SFM algorithm

under such situation might have the advantage of facing less of a local-minimum problem.

In sum, the Euclidean SFM algorithms assuming erroneous calibration parameters exhibit different behavior from the error-free case, and these deviations are more distinct when either the actual FOV or the estimated FOV is large, because then the constraint on $\hat{\alpha}$ and $\hat{\beta}$ (equation (3.13)) is stronger. As shown in Figure 3.3.1, this means that under-estimating f gives rise to more pronounced shift of the estimated FOE compared to over-estimating f (given the same magnitude in f_e). This is consistent with the somewhat paradoxical finding of [65] that larger FOV gives rise to larger bias in the translation estimate. Note, however, that over-estimating f results in a larger variance in the FOE estimate under the influence of random image noise. Equation (3.13) also means that we can recover the ratio of α to β with better accuracy. This can be seen in Figure 3.2, where with a FOV of 53^o, the curves $\frac{\hat{f}}{f}$, $\frac{\hat{\alpha}}{\alpha}$ and $\frac{\hat{\beta}}{\beta}$ increase approximately in tandem, which means that the ratio of α to β can be recovered relatively well.

3.3.2 Visualizing the Error Surface J_R

Further properties of the motion estimation process under calibration errors will be visualized through plotting the residual of the cost function J_R. Before doing so, let us discuss briefly the plotting of this surface. For easier visualization, we consider a

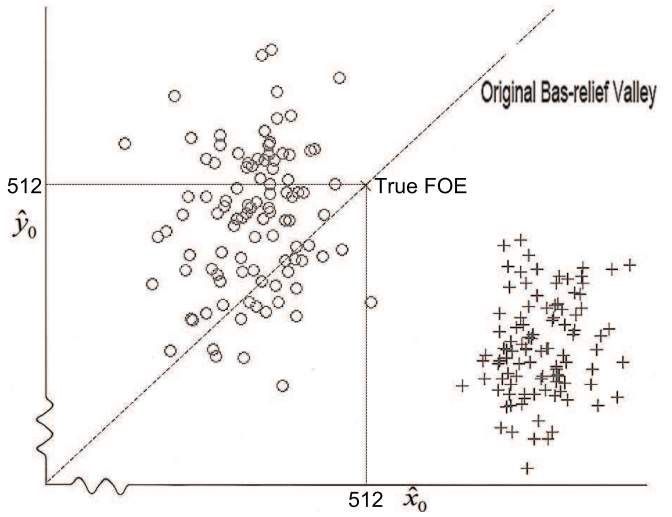


Figure 3.1: Over- and under-estimating focal length f by the same amount (i.e. same $|f_e|$) has different degree of influence on the estimation of FOE. The true FOE is marked with "×". Estimated FOEs with under- and over-estimated focal length are marked with "+" and "∘" respectively. There are 50 trials for over-estimating f and 50 trials for under-estimating f. An isotropic random noise is added to the optical flow on each trial. Under-estimating f ("+") gives rise to more pronounced shift of the estimated FOE compared to over-estimating f ("∘"); however, the latter displays a larger variance in the estimate under the influence of random image noise.

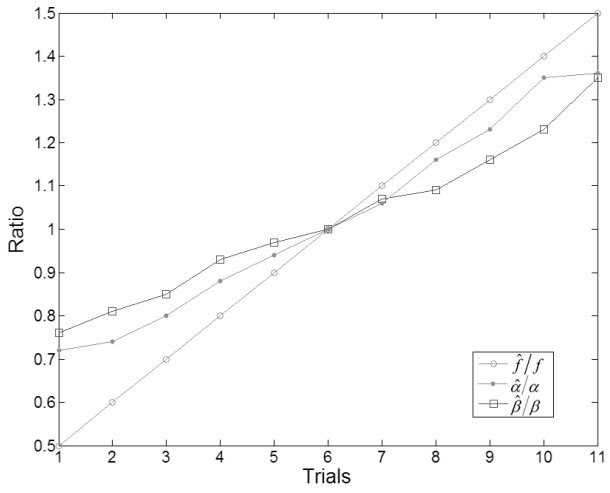


Figure 3.2: With a relatively wide FOV of 53^o, the constraint exerted on the rotational estimates $\hat{\alpha}$ and $\hat{\beta}$ is strong. The curves $\frac{\hat{f}}{f}$, $\frac{\hat{\alpha}}{\alpha}$ and $\frac{\hat{\beta}}{\beta}$ increase approximately in tandem with increasing $\hat{f}$, which means that the ratio of α to β can be recovered well.

3-dimensional surface, where each point on the surface represents a FOE hypothesis, with the height representing the residue J_R. Given a particular FOE hypothesis and a fixed (possibly erroneous) focal length estimate, the rotation variables are solved via a linear algorithm while minimizing J_R. By computing the residual error J_R for each FOE candidate, we can describe the entire residual surface completely. Some assumptions are made regarding the distribution of the feature points and the depths. We assume that the feature points are evenly distributed in the image plane, as is the distribution of the "depth-scaled feature points" $(\frac{x}{Z}, \frac{y}{Z})$. The latter assumption generally requires that the distribution of depths are independent of the corresponding image co-ordinates x and y. Different combinations of translation and rotation with over- and under-estimation of f are simulated. These simulations are carried out based on the "epipolar reconstruction" scheme, that is, setting $\mathbf{n}$ in equation (4.1) to be along the estimated epipolar direction (we reiterate that the results obtained are independent of the choice of $\mathbf{n}$). Given this scheme and for a particular FOE candidate $(\hat{(}x_0), \hat{y_0})$, J_R is given by:

$$J_R = \sum \left(\frac{c_1 - \left(c_2\hat{\alpha} + c_3\hat{\beta} + c_4\hat{\gamma} \right)}{\eta} \right)^2 \tag{3.16}$$

where

$$
\begin{aligned}
c_1 &= u(y-\hat{y_0}) - v(x-\hat{x_0}) \\
c_2 &= \frac{xy}{\hat{f}}(y-\hat{y_0}) - (\frac{y^2}{\hat{f}} + \hat{f})(x-\hat{x_0}) \\
c_3 &= \frac{xy}{\hat{f}}(x-\hat{x_0}) - (\frac{x^2}{\hat{f}} + \hat{f})(y-\hat{y_0}) \\
c_4 &= x(x-\hat{x_0}) + y(y-\hat{y_0}) \\
\eta &= \sqrt{(x-\hat{x_0})^2 + (y-\hat{y_0})^2}
\end{aligned}
$$

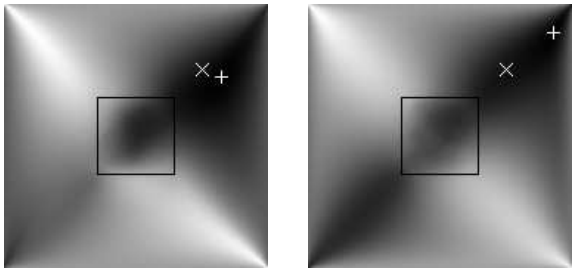
and we minimize J_R over all points in the image to solve for the rotation variables $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$. This is a typical linear least squares fitting problem, which we solved by the singular value decomposition method. We performed this fitting for each fixed FOE candidate over the whole 2-D search space and obtained the corresponding reprojected flow difference J_R. These residual values J_R were then plotted in such a way that the image intensity encoded the relative value of the residual (bright pixels corresponded to high residual values and vice versa). The imaging surface was a plane with a dimension of 512×512 pixels; its boundary was delineated by a small rectangle in the center of the plots (see Figure 3.3). The residuals were plotted over the whole FOE search space covering the entire hemisphere in front of the camera. We used visual angle in degree rather than pixel when stepping through the FOE search space; thus the coordinates in the plots were not linear in the pixel unit. The synthetic experiments have the following parameters: unless otherwise stated, the focal length was 512 pixels which meant a FOV of approximately 53^o; there were 200 feature points distributed randomly over the image plane, with depths

ranging from one to three times the focal length (i.e. 512 to 1536 pixel units). The camera was undergoing a general translation with $\mathbf{v} = (1, 1, 1)$ pixel units per second.

3.3.3 Further properties of motion estimation with calibration errors

We use the next few figures (Figures 3.3 to 3.8) to corroborate both predictions made in the preceding subsection as well as further observations made in this subsection. For all figures, true FOEs and the estimated FOEs are indicated by "$\times$" and "$+$" respectively.

1. **Influence of FOV**. Figure 3.3 illustrates the influence of visual field. Under large FOV (53^o), the second order flow field $\mathbf{t}_{2,2}$ exerts a stronger influence through equation (3.13), which constrains the value of $\hat{\alpha}$ and $\hat{\beta}$. As discussed above, this constraint on $\hat{\alpha}$ and $\hat{\beta}$ in turn reduces the length of the valley formed by the bas-relief ambiguity, while at the same time the rotation of the bas-relief valley is more pronounced, although the valley itself becomes more "diffused" and shallow (Figure 3.3a). In small FOV (28^o), the constraint (3.13) is less effective; the constraint in (3.11) can be broken down into two independent constraints like in the calibrated case, resulting in a bas-relief valley that stretches across almost the entire visual field, with little rotation

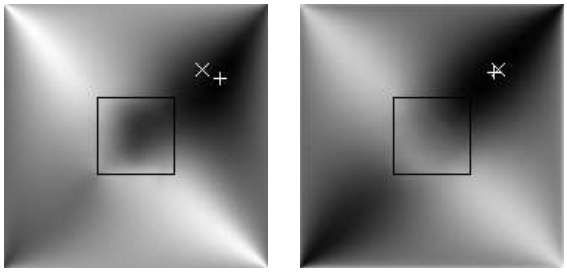


(a) (b)

Figure 3.3: The bas-relief valley is rotated if there is an error in the focal length estimate (50% under-estimated here). $\mathbf{v} = (1, 1, 1)$, $\mathbf{w} = (0.001, 0.001, 0.001)$. (a) FOV=$53^o$ (b) FOV=28^o. For all figures, true FOEs and global minima are highlighted by "×" and "+" respectively. Comparison between (a) and (b) reveals the influence of FOV on the amount of bas-relief valley rotation. Larger FOV results in larger rotation and the bas-relief valley becomes less well-defined and less elongated.

in the direction of this valley compared to the calibrated case (Figure 3.3b).

2. **Error in the estimate** $\hat{f}$. The relative importance of $\mathbf{t}_{2,2}$ is also affected by the estimated focal length $\hat{f}$. This can be seen by pitting the magnitude of the various terms of $\mathbf{t}_{2,2}$ against those of $\mathbf{t}_{2,0}$, which include among others, $\frac{\alpha}{f}$ versus αf, $\frac{\beta}{f}$ versus βf, $\frac{\hat{\alpha}}{\hat{f}}$ versus $\hat{\alpha}\hat{f}$, and $\frac{\hat{\beta}}{\hat{f}}$ versus $\hat{\beta}\hat{f}$. Given a particular f, under-estimating f (i.e. $\hat{f}$ becomes small) has the effect of enhancing the second order effect through raising $\frac{\hat{\alpha}}{\hat{f}}$ compared to $\hat{\alpha}\hat{f}$ and $\frac{\hat{\beta}}{\hat{f}}$ compared to $\hat{\beta}\hat{f}$. Thus under-estimating f would in general produce a stronger modification to the bas-relief valley compared to over-estimating f. This is clearly illustrated



(a) (b)

Figure 3.4: The influence of estimate $\hat{f}$ (with $f = 512$) on the amount of bas-relief rotation. (a)$\hat{f} = 256$, focal length under-estimated, with distinct rotation of the bas-relief valley, (b)$\hat{f} = 1024$, focal length over-estimated, but rotation of the bas-relief valley not conspicuous. Bas-relief valley also becomes less well-defined under large estimated FOV in (a).

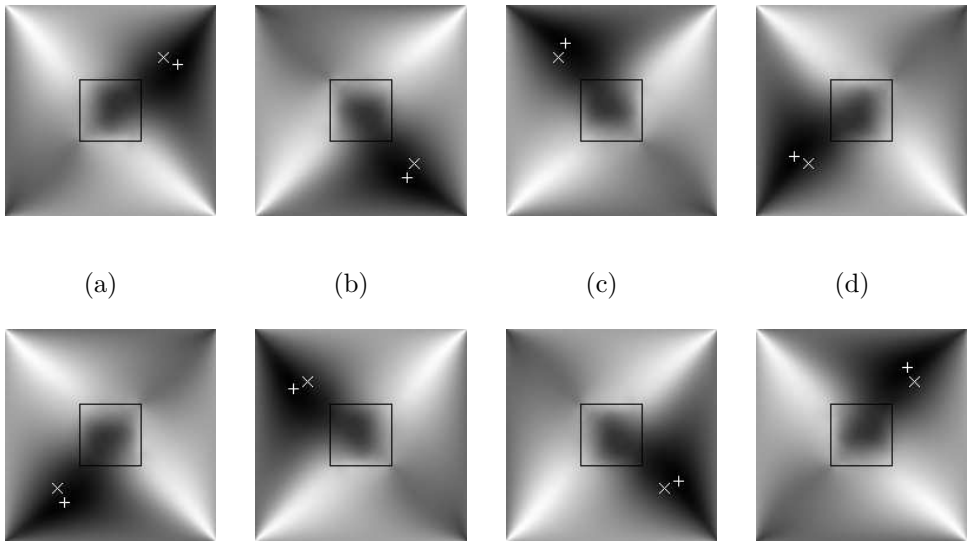
in Figure 3.4, where even the amount of f over-estimation is larger than the amount of under-estimation, the tilting of the bas-relief valley for the former (Figure 3.4b) is much less than that of the latter (Figure 3.4a). What this means is that if we want to recover the true FOE, it is better to over-estimate f than to under-estimate f. Note also that due to the larger estimated FOV in Figure 3.4a, there is a shortening of the bas-relief valley; its more diffused character is also clear.

3. **Direction of valley rotation**. Referring to equation (3.15), the direction in which the bas-relief valley rotates depends on a variety of factors such as the sign of f_e and the angle between (α, β) and (x_0, y_0). We illustrate the relationship by first looking at the case when $\alpha > 0$, $\beta > 0$, $x_0 > 0$ and

$y_0 > 0$. The direction of rotation depends on the sign of f_e in the following way. If $f_e > 0$, the signs of the terms $\alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$ and $\beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$ in equation (3.15) are both positive. It is then clear that the new slope of the bas-relief valley $\frac{\hat{y}_0}{\hat{x}_0} = \frac{y_0 - \alpha f Z\left(1-\left(\frac{\hat{f}}{f}\right)^2\right)}{x_0 + \beta f Z\left(1-\left(\frac{\hat{f}}{f}\right)^2\right)}$ deviates from the original direction $\frac{y_0}{x_0}$ (when $f_e = 0$) in a clockwise manner (Figure 3.4a). Conversely, when $f_e < 0$, the rotation in the bas-relief valley is in an anti-clockwise direction. However the amount of rotation is not so conspicuous compared to the case of $f_e > 0$ (Figure 3.4b). The reason for this anisotropy with respect to the sign of f_e has been explained earlier by their respective effects on the importance of the $\mathbf{t}_{2,2}$ term. To aid further discussion for all the other cases, we define the direction of various vectors as follows. For instance, when $\alpha > 0$ and $\beta > 0$, we say that the vector (α, β) is in the first quadrant. Carrying out the analysis for all the other cases, we find that the bas-relief valley rotates as follows. For the case of under-estimation of f, if (α, β) is in the same quadrant as (x_0, y_0), the bas-relief valley rotates in a clockwise direction (Figure 3.5, first row). Conversely, if the two vectors (α, β) and (x_0, y_0) reside in diametrically opposite quadrants, the bas-relief valley rotates in an anti-clockwise direction (Figure 3.5, second row). For the case of over-estimation of f, this relationship is exactly reversed. If the two vectors (α, β) and (x_0, y_0) are in adjacent quadrants (e.g. quadrants 1 and 2), the direction of valley rotation can be clockwise or anti-clockwise or there can be no rotation, depending on the relative magnitudes of the various terms. For instance, in Figure 3.6,

the "directions" of (x_0, y_0) and (α, β) are in the first and fourth quadrant respectively and f is under-estimated. The bas-relief valley rotates in different directions depending on the relative magnitude of α and β. If we regard the movement of the bas-relief valley as an indication of the amount of bias in the FOE estimate, caused by an error in the focal length estimate, we can see that the bias is not necessarily towards the image center but depends on a variety of factors discussed above.

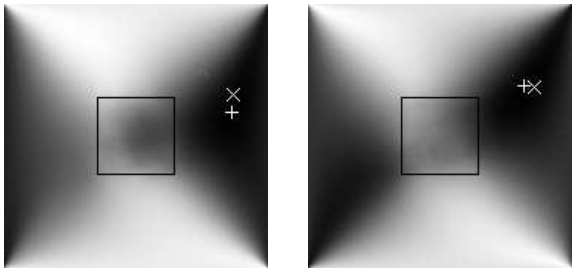
4. **Amount of FOE shift.** Having looked at the direction of the bias in the FOE estimate, we next examine the quantitative aspect of this bias, given different amount of error in the focal length estimate $\hat{f}$. Figure 3.7 illustrates the error surface for varying amount of error in the estimate $\hat{f}$, and for a relatively large FOV of 53^o under which we expect the effect of bias caused by the error in the estimate $\hat{f}$ would be more keenly felt. It can be seen that even with a rather large under-estimation error of 50% in $\hat{f}$ (the rightmost point of Figure 3.7), the relative shift in the estimate $\hat{x}_0$ is only about 37%. For the case of over-estimation in $\hat{f}$, the FOE estimate deviates very little away from the calibrated case. This anisotropy has been explained before and is due to effect of $\hat{f}$ on the relative importance of the $\mathbf{t}_{2,2}$ term, which in turn gives rise to equation (3.15). Thus, to the extent that equation (3.15) is operative, we can then characterize the maximum amount of shifts in x_0 and y_0 respectively by the two terms $\beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$ and $\alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)$ in



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

(e) (f) (g) (h)

Figure 3.5: Rotation of the bas-relief valley for (x_0, y_0) and (α, β) in different quadrants, with under-estimated focal length. In the first row, where (x_0, y_0) and (α, β) are in the same quadrant, the bas-relief valley experiences a clockwise rotation; whereas in the second row, where (x_0, y_0) and (α, β) are in diametrically opposite quadrants, the bas-relief valley rotates in an anti-clockwise direction. $W = 1$, $\gamma = 0.001$ $f = 512$ and $\hat{f} = 256$ for all figures. The (U, V) and (α, β) are respectively (a) $(1, 1)$, $(0.001, 0.001)$ (b) $(1, -1)$, $(0.001, -0.001)$ (c) $(-1, 1)$, $(-0.001, 0.001)$ (d) $(-1, -1)$, $(-0.001, -0.001)$ (e) $(-1, -1)$, $(0.001, 0.001)$ (f) $(-1, 1)$, $(0.001, -0.001)$ (g) $(1, -1)$, $(-0.001, 0.001)$ (h) $(1, 1)$, $(-0.001, -0.001)$.



(a) (b)

Figure 3.6: Rotation of bas-relief valley when the "directions" of (x_0, y_0) and (α, β) are in adjacent quadrants. $(U, V, W) = (3, 1, 1)$, $f = 512$, and $\hat{f} = 256$. Residual error maps are plotted with (a) $(\alpha, \beta, \gamma) = (0.003, -0.001, 0)$, and (b) $(\alpha, \beta, \gamma) = (0.001, -0.007, 0)$. The direction of rotation is clockwise for (a) and anti-clockwise for (b).

that equation. To pin down the value for such a bound, we assume that the effect of Z in the above two terms can be represented by some average depth Z_{ave}. Then in relative terms, the changes to x_0 can be expressed as follows:

$$\begin{aligned} \frac{x_0 - \left(x_0 + \beta f Z_{ave}\left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)\right)}{x_0} &= \frac{\beta f}{fU/Z_{ave}} W\left(1 - \left(\frac{\hat{f}}{f}\right)^2\right) \\ &\approx \frac{u_{pan}}{u_{trans-x}} W\left(1 - \left(\frac{\hat{f}}{f}\right)^2\right) \end{aligned} \quad (3.17)$$

where u_{pan} and $u_{trans-x}$ are respectively the horizontal flow components due to panning rotation β and lateral translation U with some average depth Z_{ave}. Similar expression can be obtained for the relative change in the estimate for y_0. It can be seen that the relative change is affected by the ratio of the rotational flow u_{pan} and the translational flow $u_{trans-x}$; which is in turn

moderated by a multiplicative factor $W\left(1-\left(\frac{\hat{f}}{f}\right)^2\right)$. Thus, for the simulation conducted in Figure 3.7, where the translational flow and rotational flow are approximately equal in magnitude and $W = 1$, a large under-estimation error of 50% in $\hat{f}$ would result in a bound of 75% in the FOE shift. That this bound is much larger than the actual shift (37%) obtained could be due to violation of the two assumptions made in deriving this bound: (1) the $\mathbf{t}_{2,2}$ term is maximally effective, and (2) scene points at different depths play an equal role such that their effect can be represented by some average depth Z_{ave}. Despite the looseness and approximate nature of the bound, we can use equation (3.17) as a guide in assessing whether the resulting bias in FOE is acceptable when using an approximate value of the focal length in a calibrated SFM algorithm, or it is better to face the tricky problem of estimating the focal length (as discussed in [5, 42, 49]) using a general uncalibrated SFM algorithm. As an illustrative example, consider a more typical error of 10% in the estimate $\hat{f}$ and under the same motion-scene configuration as above: the bound obtained via equation (3.17) for the relative FOE shift would be 19% (for under-estimation of f). Furthermore, this is likely to be a very loose bound; the actual shift obtained in the simulation is only 4%. Thus we might want to proceed with a calibrated SFM algorithm even though the focal length estimate has small error.

5. **Effect of erroneous principal point.** Besides being affected by error in

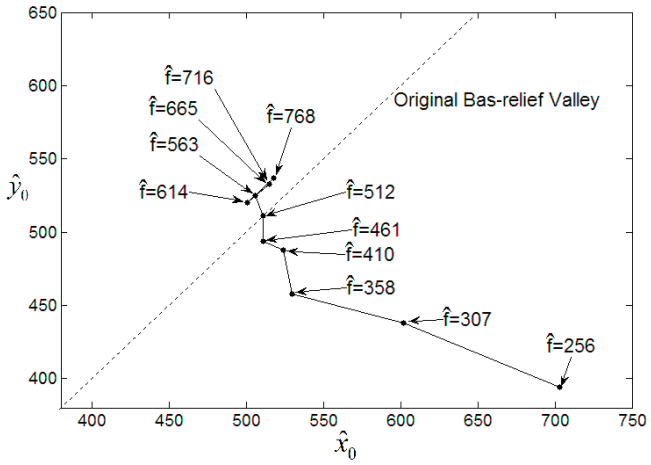


Figure 3.7: The amount of shift in the estimated FOE with different errors in the estimated focal length. The true focal length is 512, whereas the estimated focal length vary from 256 (50% under-estimation) to 768 (50% over-estimation), with a step size of 10% error. The translational and rotational parameters are $(U, V, W) = (1, 1, 1)$ and $(\alpha, \beta, \gamma) = (0.001, 0.001, 0.001)$ respectively. True FOE lies at the point $(512, 512)$ on the bas-relief valley. The estimated FOEs deviate very little away from the true solution for the case of over-estimation in $\hat{f}$. For the case of under-estimation in $\hat{f}$, the amount of shift in the FOE is more significant. However, even with a rather large under-estimation error of 50% in $\hat{f}$, the relative shift in the estimate $\hat{x}_0$ is only about 37%.

the focal length estimate, the bas-relief valley is also changed by error in the principal point estimate. We use (x_s, y_s) to represent an image pixel location in an image coordinate system with its origin located at the lower left corner of the image. If the principal point of the camera is situated at (O_x, O_y) in this new coordinate system, then (x, y) and (x_s, y_s) are related by $(x, y) = (x_s - O_x, y_s - O_y)$. Given an error (O_{x_e}, O_{y_e}) in the principal point estimate, the corresponding error function J_R can be shown to be given by[2]

$$J_R = \sum \left(\frac{(x + O_{x_e} - \hat{x}_0, y + O_{y_e} - \hat{y}_0) \cdot \left(v_{rot_e} - \frac{y_{0_e} + O_{y_e}}{Z}, \frac{x_{o_e} + O_{x_e}}{Z} - u_{rot_e} \right)}{(x + O_{x_e} - \hat{x}_0, y + O_{y_e} - \hat{y}_0) \cdot \mathbf{n}} \right)^2$$

where u_{rot_e} and v_{rot_e} are given by:

$$\begin{aligned} u_{rot_e} &= -\left(\beta f - \hat{\beta}\hat{f}\right) + \frac{\alpha}{f}xy - \frac{\hat{\alpha}}{\hat{f}}(x + O_{x_e})(y + O_{y_e}) \qquad (3.18) \\ &\quad -\frac{\beta}{f}x^2 + \frac{\hat{\beta}}{\hat{f}}(x + O_{x_e})^2 + \gamma y - \hat{\gamma}(y + O_{y_e}) \\ v_{rot_e} &= \left(\alpha f - \hat{\alpha}\hat{f}\right) + \frac{\alpha}{f}y^2 - \frac{\hat{\alpha}}{\hat{f}}(y + O_{y_e})^2 - \frac{\beta}{f}xy \qquad (3.19) \\ &\quad +\frac{\hat{\beta}}{\hat{f}}(x + O_{x_e})(y + O_{y_e}) - \gamma x + \hat{\gamma}(x + O_{x_e}) \end{aligned}$$

[2]Note that in deriving these equations and plotting the figures, the true and the estimated FOEs should be independent of the choice of the principal point, as the FOE actually indicates a direction in space—that of the 3D translation.

The corresponding terms in $\mathbf{t}_1$ and $\mathbf{t}_2$ are:

$$\begin{aligned}
\mathbf{t}_{1,0} &= (-\hat{x}_0 + O_{x_e}, -\hat{y}_0 + O_{y_e})^T \\
\mathbf{t}_{1,1} &= (x, y)^T \\
\mathbf{t}_{2,0} &= \left(\left(\alpha f - \hat{\alpha}\hat{f}\right), \left(\beta f - \hat{\beta}\hat{f}\right)\right)^T \\
\mathbf{t}_{2,1} &= \left(-\gamma x + \hat{\gamma}\left(x + O_{x_e}\right), -\gamma y + \hat{\gamma}\left(y + O_{y_e}\right)\right)^T \\
\mathbf{t}_{2,2} &= \left(\frac{\alpha}{f}y^2 - \frac{\hat{\alpha}}{\hat{f}}\left(y + O_{y_e}\right)^2 - \frac{\beta}{f}xy + \frac{\hat{\beta}}{\hat{f}}\left(x + O_{x_e}\right)\left(y + O_{y_e}\right),\right. \\
&\quad \left. -\frac{\alpha}{f}xy + \frac{\hat{\alpha}}{\hat{f}}\left(x + O_{x_e}\right)\left(y + O_{y_e}\right) + \frac{\beta}{f}x^2 - \frac{\hat{\beta}}{\hat{f}}\left(x + O_{x_e}\right)^2\right)^T \\
\mathbf{t}_{2,Z} &= \left(-\frac{y_{0_e} + O_{y_e}}{Z}, \frac{x_{0_e} + O_{x_e}}{Z}\right)^T
\end{aligned}$$

To derive the conditions conducive for the formation of the bas-relief ambiguity, we apply the same condition that the constant-direction vectors $(\mathbf{t}_{2,0} + \mathbf{t}_{2,Z})$ and $\mathbf{t}_{1,0}$ should be perpendicular to each other. We obtain, analogous to equation (3.11), the following:

$$\frac{y_{0_e} + O_{y_e} - \alpha f Z + \hat{\alpha}\hat{f}Z}{x_{0_e} + O_{x_e} + \beta f Z - \hat{\beta}\hat{f}Z} = \frac{\hat{y}_0 - O_{y_e}}{\hat{x}_0 - O_{x_e}} \tag{3.20}$$

The corresponding condition for making $||\mathbf{t}_{2,2}||$ small gives rise to the following:

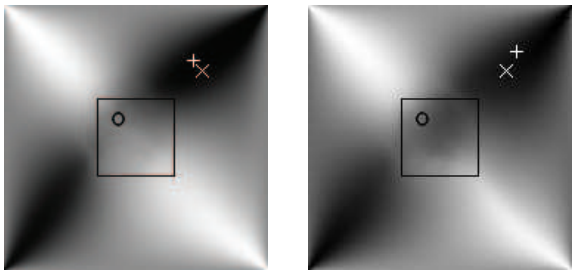
$$\begin{aligned}
\frac{\alpha xy}{f} &= \frac{\hat{\alpha}(x + O_{x_e})(y + O_{y_e})}{\hat{f}} \\
\frac{\beta xy}{f} &= \frac{\hat{\beta}(x + O_{x_e})(y + O_{y_e})}{\hat{f}} \\
\frac{\alpha y^2}{f} &= \frac{\hat{\alpha}(y + O_{y_e})^2}{\hat{f}} \\
\frac{\beta x^2}{f} &= \frac{\hat{\beta}(x + O_{x_e})^2}{\hat{f}}
\end{aligned}$$

which are obviously not satisfiable at all points of the image. However, if we make the assumption that the second order effect $||\mathbf{t}_{2,2}||$ only comes into play at the peripheral image points where x and y are large and that the magnitude of the error (O_{x_e}, O_{y_e}) is small compared to x and y at these peripheral points, then the original constraint $\frac{\alpha}{f} = \frac{\hat{\alpha}}{\hat{f}}$, $\frac{\beta}{f} = \frac{\hat{\beta}}{\hat{f}}$ of equations(3.13) is still approximately true. Substituting this into equation (3.20), we obtain, after some manipulation, the following form:

$$\frac{y_{0e} - \alpha f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)}{x_{0e} + \beta f Z \left(1 - \left(\frac{\hat{f}}{f}\right)^2\right)} = \frac{\hat{y}_0 - O_{y_e}}{\hat{x}_0 - O_{x_e}}$$

which is analogous to equation (3.14). It differs from equation (3.14) in that the bas-relief valley has been translated by an uniform amount (O_{x_e}, O_{y_e}) and passes through the true principal point. Figure (3.8) illustrates the changes caused by $(O_{x_e}, O_{y_e}) = (100, -100)$, for (a) when there is no error in $\hat{f}$, and (b) when there is an under-estimation error of 50%. The bas-relief valleys appear bent because we have used visual angle in degree rather than pixel as the FOE search step and thus the co-ordinates in the plots were not linear in the pixel unit.

6. **Implication for various visual tasks**. We have seen how the recovery of the FOE is affected by errors in the calibration parameters. How do these errors affect metric depth recovery? In [12], we have shown that the type of motion executed is crucial for depth recovery. Under lateral movement,



(a) (b)

Figure 3.8: The bas-relief valley with erroneous principal point estimate $(\hat{O}_x, \hat{O}_y) = (0, 0)$. The entire bas-relief valley is shifted by a constant amount and passes through the true principal point at (100, -100) (indicated by "○"). The bas-relief valleys appear bent because we have used visual angle in degree rather than pixel as the FOE search step and thus the co-ordinates in the plots were not linear in the pixel unit. $(U, V, W) = (3, 1, 1)$, $(\alpha, \beta, \gamma) = (0.003, -0.001, 0)$, and $f = 512$. (a) $\hat{f} = 512$ (b) $\hat{f} = 256$ (50% under-estimation).

while it might be very difficult to resolve the ambiguity between translation and rotation, depth orders of scene points can be recovered with robustness. Conversely, under forward translation, it is difficult to recover structure unless favorable conditions such as large field of view exist, because under this motion configuration, small error in the FOE estimate can introduce large distortion in the depth recovered. In the case of uncalibrated motion, in spite of uncertainty in the focal length, the qualitative aspect of the depth recovery process is not affected, regardless of whether it is a lateral or a forward motion. That is, under lateral motion, despite possible rotation of the bas-relief valley, the depth orders of scene points are shown in [12] to be preserved. Conversely, under forward motion, the inherent difficulty in depth recovery would have been compounded by the errors in the intrinsic parameters, as we have shown earlier that errors in the intrinsic parameters introduce additional bias to the FOE estimate.

Let us explore the ecological implications even we do suffer from depth distortion when we are executing forward motions. Such motions are mainly used in moving towards an object or for navigating through an environment. In the context of such tasks, we might only need aspects of structural information to successfully complete the tasks, rather than acquiring a comprehensive metric scene reconstruction. For instance, the ability to estimate the time-to-collision (TTC) is important for avoiding collision. It has been

argued [57, 70, 93] that TTC can be recovered directly from the first order derivatives of the optical flow, without going through the step of 3D motion recovery. As a consequence, the TTC estimate would not be affected by the aforementioned depth distortion, which stems from errors in the 3D motion recovery. Nevertheless, calibration errors do affect the TTC estimate even it is recovered directly from the optical flow. In the calibrated case, the TTC estimate is not exact but bounded by some deformation terms [93] depending on the amount of lateral translation and the surface slant. If there now exists some error in the principal point estimate, the TTC bound would be affected by this error. The detailed examination of how such task-specific structural information is affected by calibration errors, while interesting, is beyond the scope of this thesis.

Another commonly encountered scenario is that of a motion fixating on a point of an object. One example of such scenario is a camera rotating around an object in image-based modelling or image-based rendering application. If we assume that the fixation is accomplished via the pan and tilt rotation, as is usually done, then it can be easily shown that (x_0, y_0) and (α, β) are always in adjacent quadrants with $\frac{y_0}{x_0} = -\frac{\alpha}{\beta}$. The fixation constraint also allows us to show that the signs and magnitudes of α and β would be such that the bias of the FOE estimate is always towards the optical center, along the direction of the original bas-relief valley. Note that this is also the condition under which

[65] carried out their simulations and obtained the results that the bias in the FOE estimate caused by the error in $\hat{f}$ is towards the optical center. Our model confirms this result under this specific motion configuration but also predicts other bias directions under more general motion configurations.

3.4 Experiments and discussion

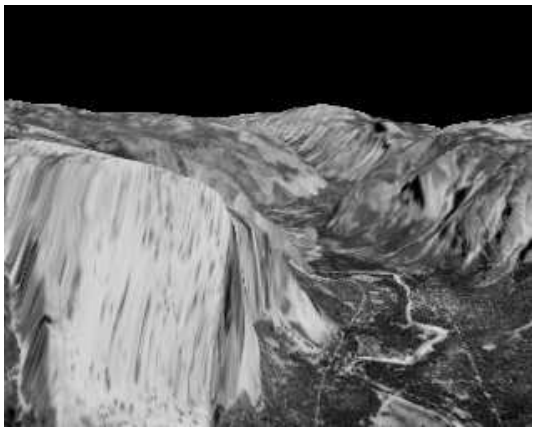
To verify the theoretical findings established just set out, we perform experiments on both the Yosemite sequence and the Coke sequence. The optical flow was obtained using Lucas-Kanade algorithm [62] with a temporal window of 11 frames. Relatively dense optical flow fields were obtained. The cost function was implemented based on the "epipolar reconstruction" scheme, that is, setting $\mathbf{n}$ in equation (4.1) to be along the estimated epipolar direction. We demonstrate that given fairly dense and uniform distribution of scene points, our predictions about the changes to the bas-relief valley and the bias in the FOE estimate due to erroneous focal length hold true.

In the first experiment, the computer generated Yosemite sequence (Figure 3.9a) was used. The average FOV is 46^o, the true focal length is 337.5 pixels, and the true FOE is located at $(0, 59.5)$. Figure 3.9b shows the estimated FOE locations for $\hat{f}$ having errors of $0\%, \pm 16\%, \pm 33\%$, and $\pm 50\%$.

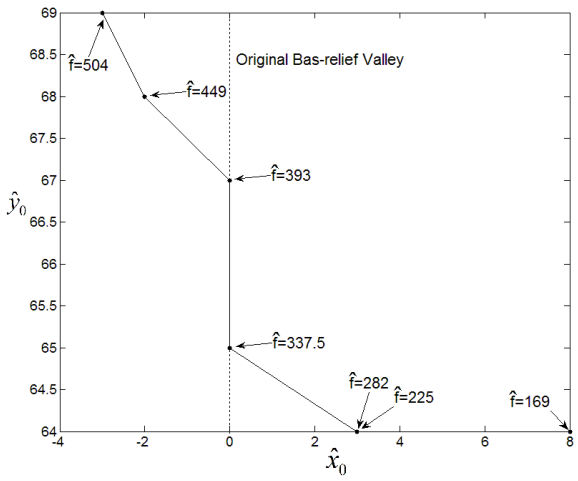
In the second experiment, similar analysis was conducted on the Coke image sequence (Figure 3.10a). The parameters of this sequence are FOV=28^o, f=620 pixels, and the true FOE at $(65, 73)$. The experimental results are shown in Figure 3.10b.

The results obtained seem to corroborate the various predictions made in this thesis. In both sequences, the direction of bias in the FOE estimate is consistent with the predictions made in the preceding section. The actual amount of FOE shift is also unanimously small even for a large error of 50% in $\hat{f}$, the shifts being less than 10 pixels in both cases. We also predicted that the bias will be less pronounced for over-estimating rather than under-estimating f, though this prediction is not borne out by the results, with the case of over-estimation exhibiting comparable amount of FOE shift as that of under-estimation for both sequences. However, this is not surprising as we can see from Figure 3.3.1 that in the case of over-estimation, the FOE estimate, while displaying a smaller bias, suffers from a larger variance under the influence of noise. With the significant effect of local minima introduced by non-uniform feature distribution and the presence of noise in real images, this high variance term becomes important, thus contributing to the larger-than-expected FOE errors seen in the results. In fact, as can be seen from Figures 3.9b and 3.10b, these non-ideal effects also hamper the FOE recovery under perfect calibration, with the direction of FOE errors lying along the bas-relief valley.

Overall, we found that the actual shift in the FOE estimate for real images is



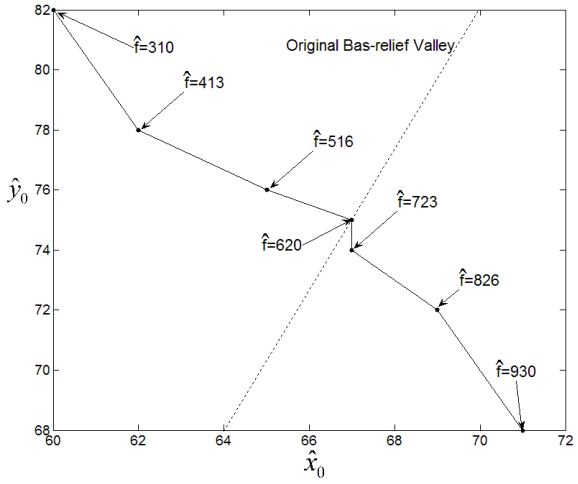
(a)



(b)

Figure 3.9: (a) Yosemite sequence. (b) Shift of the FOE estimate as a result of erroneous focal length estimate $\hat{f}$. The true focal length of the image sequence is 337.5 and the true FOE is at $(0, 59.5)$. Estimated FOEs are plotted for $\hat{f}$ having errors of $0\%, \pm 16\%, \pm 33\%$, and $\pm 50\%$ respectively.

(a)



(b)

Figure 3.10: (a) Coke sequence. (b) Shift of the FOE estimate as a result of erroneous focal length estimate $\hat{f}$. The true focal length of the image sequence is 620 and the true FOE is at $(65, 73)$. Estimated FOEs are plotted for $\hat{f}$ having errors of $0\%, \pm 16\%, \pm 33\%$, and $\pm 50\%$ respectively.

not significant even for relatively large error in the focal length estimate. The two experiments conducted demonstrate that, even with a relatively dense set of feature points, non-ideal effects such as non-uniform feature distribution and image noise, rather than calibration errors, could play a potentially more significant role in affecting the accuracy of FOE recovery. For image sequence where the feature points are very clustered and sparse, or when the scene depths are near to a planar scene, there can be a significant change in the bas-relief ambiguity, as detailed in [110].

3.5 Conclusions

Error analysis for SFM has always been plagued by the complexity of the problem. This complexity becomes even more daunting in the face of possible calibration errors. In this chapter we have developed clear analytical expressions describing the error behavior of the egomotion estimates when the fixed intrinsic parameters are calibrated with error. The key results in this chapter are independent of the algorithm used to perform egomotion estimation and calibration. As a result of error in the estimate $\hat{f}$, the bas-relief valley is rotated in a direction that depends on the relationship between the translation and the rotation. Under-estimating the focal length would have the effect of shortening the bas-relief valley and making it less well-defined in character. It also gives rise to a larger bias in the FOE

estimate though with a smaller variance. On the other hand, over-estimating the focal length results in less change to the bas-relief valley and the FOE estimate would have smaller bias but larger variance. We also obtain an analytical bound that quantifies the effect of an erroneous focal length on the FOE estimate. For a typical figure of 10% error in the estimate $\hat{f}$ and given certain generic motion-scene conditions (such as rotation not too dominant), the bound obtained for the relative FOE shift might turn out to be acceptable. Furthermore, this bound is likely to be conservative as the actual shift obtained in simulation is consistently much smaller. Error in the principal point estimate is shown to result in a simple change to the error surface. The entire bas-relief valley is shifted by a constant amount such that it passes through the true principal point. Real-world effects such as image noise and non-uniform feature distribution are briefly investigated in the experimental section, with results showing that these non-ideal effects are likely to play a much more significant role than the errors in the calibration parameters.

The conclusion of this chapter is that if the image quality is acceptable and the feature distribution is relatively dense and uniform, we might want to use a calibrated SFM algorithm even though the focal length estimate or the principal point estimate has small errors. The resultant small loss in accuracy might be acceptable compared to the uncertainty faced in estimating the focal length or principal point using a general uncalibrated SFM algorithm. If, however, one has to deal with high image noise or sparse and clustered feature distribution, the perennial

problems that plague SFM estimation even for the calibrated case would certainly be compounded by the calibration errors, posing grim problems for any general 2-frame SFM recovery algorithm. One suspects that under these situations (such as in the real world), the visual system has to press maximal benefit from the opportunities afforded by bodily and environmental resources along with significant coupling of perception and action in order to carry out visuo- motor tasks successfully.

Chapter 4

What We See In the Cinema: A Dynamic Account

Cinema viewed from a location other than a Canonical Viewing Point (CVP) presents distortions to the viewer in both its static and dynamic aspects. Past works have investigated mainly the static aspect of this problem and attempted to explain why viewers still seem to perceive the scene very well. The dynamic aspect of depth perception, which is known as structure from motion, and its possible distortion, have not been well investigated. In our work, we derive the dynamic depth cues perceived by the viewer and use the so-called iso-distortion framework to understand its distortion. The result is that viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems.

Despite this shift, the key properties of the perceived depths remain largely the same, being determined in the main by the accuracy to which extrinsic motion parameters can be recovered. For a viewer seated at a non-central position and watching the movie screen with a slant angle, the view is related to the view at the CVP by a homography, resulting in various aberrations such as non-central projection.

4.1 Problem statements

Three projections underlie the creating and viewing of motion pictures, namely, (a) the projection from the 3-D real scene to the film of the camera, (b) the back projection from the film onto the viewing screen and (c) the projection from the screen to the human retina. These projections are assumed to be perspective in this thesis.

Mathematically, only the audience located at a certain viewing position sees a "veridical" version of the scene as if he or she is seeing through the directors eyes and making the same movement. We call this position the canonical viewing position (CVP). All other positions receive visual stimuli different from the veridical version; the differences include dynamic visual cues such as optical flow, as well as depth information arising from such dynamic cues. Paradoxically, picture viewing

is apparently not limited to the location at the CVP. Remarkably large number of positions in front of the projector can serve as reasonable viewpoints allowing layout within the motion picture to appear relatively normal. It is fortunate that the human visual system has this ability, for without it, the design of cinema theater and home entertainment system would be severely constrained.

The paradox of the unnoticed distortions was studied by researchers for about two decades. Cutting [21] argues that the slant at which pictures are viewed is usually small, and consequently the distortions of the retinal image are too small to be noticed. Perkins [77] claims that such invariance is a byproduct of the viewer's expectations with known shapes. For example, if the retinal image is similar to the image that would be created by a cube, prior expectations force the percept to that of a cube. The invariance thus comes from the viewer's experience with object whose shapes are familiar or usually follow certain rules (right angles, parallel sides, symmetry). A third explanation claims that the invariance is the consequence of altering or re-interpreting the retinal image by recovering the position of the screen surface. For example, it is known [8] that the vanishing points of three mutually orthogonal lines are sufficient to recover the principal point. Banks et al. [3] argues that a local slant mechanism is used to estimate the foreshortening due to viewing obliqueness and then adjust the percept derived from the retinal image to undo the foreshortening.

All these hypotheses mainly attempt to deal with the static aspect of the paradox.

Yet, cinema is very much an art of camera motion, as testified by the original names of kinetoscope and moving pictures. For Metz [16] indeed, movement is the principal reason for the effect of reality within film. Motion dynamically changes the viewing perspectives of the spectators both in space and in time to give the unique reality effect, allowing the viewers to inhabit the visual space of the person(s) producing the film narrative. The depth information carried by motion cues is particularly relevant as cinema is typically viewed from a distance of 20m or more, condition under which accommodation, convergence, and stereoscopic depth perception are inactive. Last but not least, it is often through motion that the content or the meaning in a shot is expressed and the attention of the viewers captivated or shifted, allowing the films intentions to be communicated. Thus motion cue and depth perception arising from it should be the privileged object of investigation in cinematic perception.

In theory, the optical flow present in the motion pictures and the dynamic depth cues arising thereof should also experience distortion but have received very little attention. In fact, it is not even clear what sort of distortion is experienced by the viewer as far as the dynamic aspect is concerned. This neglect is partly due to the fact that the distortion of depths arising from errors in the motion cues is an analytically complex problem; geometrical analysis which shed light on this problem has only been recently formulated [10, 12]. Our work focuses on this dynamic aspect of cinematic perception and investigates computationally the distortion of

both the camera motion parameters and the depth recovered from such distorted motion cues.

To recover the spatial structure using optical flow present on the picture screen amounts to the classical structure from motion (SFM) problem with a slight twist. We will introduce this modified SFM model in Section 4.3. The typical SFM problem has been the central problem of computer vision since 1980s. It recovers the structure of 3-D scene and the 3-D relative motion between the scene and the observer from the projection of the 3-D relative motion onto a 2-D surface. If the 3-D motion parameters can be estimated perfectly, depth recovery can be achieved accurately; in other words, one can perceive the spatial arrangement of objects. However, this veridical space recovery from SFM is difficult to achieve, as has been shown both computationally and experimentally. Either the 3-D motion estimates contain errors with the result that depths are distorted; or the intrinsic parameters of the camera are unknown, in which case one can only recover the so-called projective depth[106], which is related to the true depth by a projective transformation.

Since errors in motion estimates are highly likely, there have been various error analyses in the past [1, 24, 38, 76, 95, 111], in terms of the local minima and ambiguities of the SFM algorithms. However, there is much less analyses on the behaviour of depth distortion given some errors in the motion estimates. Cheong et al. [10] developed a geometric account of the depth error behaviour via the so-called

iso-distortion framework. It showed that even with known intrinsic parameters but with errors in the 3-D motion estimates, the distortion transformation from physical to perceived space is already highly complex, in fact, more complicated than that of the projective transformation. It is a space Cremona transformation which is a rational transformation between two projective spaces [48]. Given such potentially complex distortion behaviour, Cheong and Xiang [12] then motivated the importance of special generic motions favored by biological visual systems. One such motion is the lateral motion which consists of lateral translation plus rotation. Such motion will, despite errors in the estimates, yield a special type of Cremona transformation that preserves depth order. We say that such transformation exhibits ordinal depth invariance. Another generic motion type is the forward motion (forward translation plus rotation) which gives rise to conditions conducive for 3-D motion recovery but not for depth recovery. The idea here is that different motion types are suited for specific tasks; this is important since there is no general motion algorithm that can work well under all motion-scene configurations.

The SFM process is further complicated by the presence of intrinsic parameters such as the focal length and the principal point coordinates. Cheong and Xiang [12] further showed that as long as the focal length is not dynamically varying (i.e. the camera is not performing zoom operation) and the errors in the principal point estimation is small enough, the aforementioned properties of spatial perception under different generic motions are still preserved.

Whether visual systems in nature have a precise knowledge of the eyes' intrinsic parameters when processing visual tasks is still unknown. Nevertheless psychophysics researchers studying the perception of the scene structure from dynamic cues [19, 98, 25] tend to assume that the brain uses a calibrated visual system and neglect the problem of calibration altogether. This is mainly due to the elaborate model needed to describe the complex intrinsic parameters of human eyes, making it very difficult to incorporate them into computational analysis. In this thesis, we only consider the typical intrinsic parameters used for modeling pinhole camera [28]. The extent to which these intrinsic parameters are calibrated determines the type of space that can be perceived from motion cues. As to the geometric structure of this perceived visual space, there has been a host of models being proposed, e.g., Euclidean geometry [34], hyperbolic [63], affine [97], and others [47, 52]. Recently Droulez and Cornilleau-Pérès' anamorphosis glasses [26] show that the visual system is able to re-calibrate a Riemannian metric adapted to the glasses deformation and an Euclidean geometry can be perceived after the plastic adaptation. Other experimental result [69] which supports the assumption that brain cognition is more "Euclidean than affine or projective" is that when perceiving the orientation of a surface drawn using curves, subjects preferentially consider the orthogonality cue rather than parallelism. Viéville et al. [105] report that the human visual system is able to take intrinsic parameter variations into account during perceptual tasks. It has also been argued [37] that the more recently evolved vision-for-perception system is quite different from those of the more ancient vision-for-action system,

and the latter is based on Euclidean object metrics. In spite of these results, there are psychophysical evidences that suggest human vision is not Euclidean under all conditions [13, 96], especially in the impoverished scenarios typically encountered in psychophysical experiments (e.g. random dots in motion). For instance Cheong et al. [13] reported that the recovery of curvatures under lateral translation is subject to varying degrees of uncertainty depending on the motion-scene configuration. In particular, the theory proposed therein explained why the reconstructed second order shape tends to be more distorted in the direction parallel to the translational motion than that in the orthogonal direction. This orientational anisotropy has also been reported in many psychophysics papers [18, 72, 83]. [23, 22] studied the perception of second order shapes under active vision, and it was found that some types of shapes can be perceived quite accurately, whereas others are more difficult to be distinguished. Thus, on the whole, it seems that human vision is quite plastic and grades from being nearly Euclidean to non-metrical depending on tasks and conditions.

In our work, we seek to use the iso-distortion framework to analyze the nature of the depths recovered from dynamic cues under the cinema configuration. We show that viewing a movie in a cinema from a general position differs from viewing a 3-D real scene primarily in that the visual system experiences an altered optical flow resulting from changed intrinsic parameters. To be exact, this statement is only correct for a viewer seated at a reasonably central position. The impacts

on depth perception from different seating positions are elucidated and compared with SFM under normal condition. Results show that even with the shift in the intrinsic parameters, the key properties of the recovered depth remain largely the same, despite some differences from the case of normal uncalibrated SFM discussed by Cheong and Xiang [12], and these key properties are determined primarily by the degree of accuracy to which the extrinsic parameters can be recovered and by the types of motions being executed. In sum, the main contribution of our work is to show the geometric laws governing distortions in the perceived space and to make explicit those situations that lead to different types of distortions. The implications of these results for cinematic viewing and uncalibrated vision in general will be further discussed later, but these speculative possibilities have to be further investigated by comprehensive psychophysical tests, in the light of the types of distortion and their motion-scene dependency that are unraveled here.

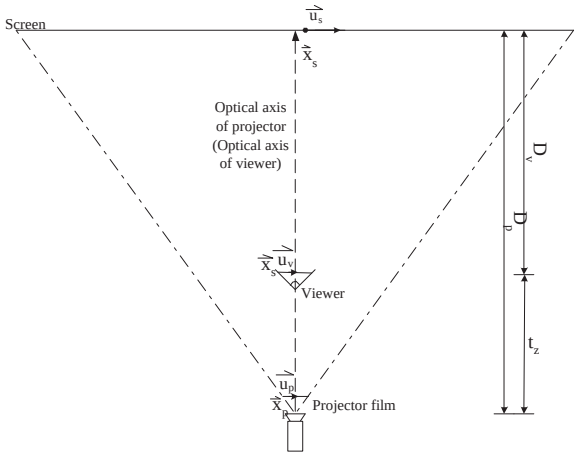
4.2 Model and Prerequisite

If the 3-D motions have been estimated, Z can be in turn obtained from equation (2.5) or (2.6). Usually a direction $\mathbf{n}$ is chosen according to some criteria to recover Z. Thus Z can be obtained as

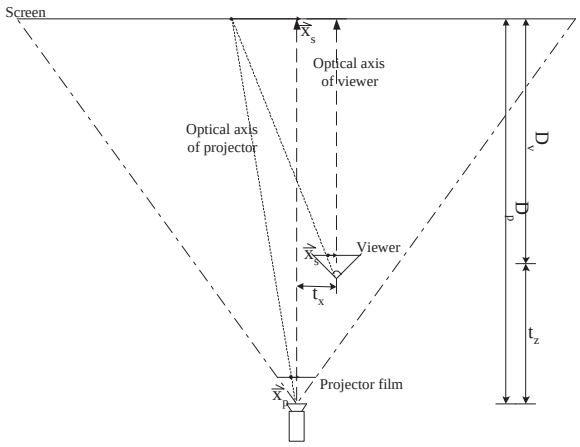
$$Z = \frac{(x - x_0, y - y_0) \cdot \mathbf{n}}{(u - u_{rot}, v - v_{rot}) \cdot \mathbf{n}}. \tag{4.1}$$

In the above equation, $(x_0, y_0) = \left(f\frac{U}{W}, f\frac{V}{W}\right)$ denotes the focus of expansion (FOE), (u_{rot}, v_{rot}) are the rotational components of equations (2.5) and (2.6) respectively, and $\mathbf{n}$ is the unit vector in the direction chosen to recover Z. As an example, $\mathbf{n}$ can be along the normal flow direction because the flow along this direction can be most reliably estimated. In the case where optical flow can be recovered well, other considerations might lead to the choice of recovering depth along the direction emanating from the estimated FOE (x_0, y_0), based on the intuition that this direction (also known as the epipolar direction) contains the strongest translational flow and thus provides the best estimate of depth.

It follows that if there are some errors in the estimation of the extrinsic parameters, Z will be estimated with errors, that is, a distorted version of the space will be perceived. The detailed analysis of this depth distortion will be deferred to Section 4, after we have introduced the modified form of the SFM problem under the cinema viewing configuration.



(a)



(b)

Figure 4.1: A simple cinema viewing configuration. $\vec{x}_p$, $\vec{x}_s$ and $\vec{x}_v$ represent respectively the feature points on the projector film, screen, and viewer's retina corresponding to the same world point. (a) optical axes of viewer and projector are coincident (b) optical axes of viewer and projector are not coincident but parallel to each other.

4.3 Structure from motion under cinema viewing configuration

4.3.1 Optical axes of viewer and projector parallel

We first consider the case whereby the viewer's optical axis is parallel to the projector's optical axis, and the screen is oriented in a fronto-parallel manner to the projector and the viewer. This is applicable to most cinema viewers who are seated not near the side or right at the front (Figure 4.1). As the seats are designed to face forward, the viewers will do so unless they are positioned so far off that they are obliged to tilt their viewing axis towards the central area of the screen. We assume the cinema images captured by the director has been transferred to film for optical projection and we call this film the projector film. We also assume monocular viewing to focus on just motion cue. We use subscripts $p,\ v,\ s$ to represent quantities associated with projector, actual viewer and screen, respectively. The distances (along the Z-axis) from the screen to the projector and to the viewer are D_p and D_v, respectively. The focal length of the projector and that of the viewer's visual system are f_p and f_v, respectively.

Consider the simplest case where the viewer's optical axis is not only parallel to but also coincident with the projector's optical axis. Then clearly the feature points

$\vec{x}_p$, $\vec{x}_s$ and $\vec{x}_v$ (see Figure 4.1(a)) are related by:

$$\vec{x}_p = \frac{f_p \vec{x}_s}{D_p} \tag{4.2}$$

$$\vec{x}_v = \frac{f_v \vec{x}_s}{D_v} = \frac{\vec{x}_p}{k}, \tag{4.3}$$

where $\vec{x}_s$ is given in metric unit, $\vec{x}_p$ and $\vec{x}_v$ are in pixel units and

$$k = \frac{D_v f_p}{D_p f_v}$$

Assume there is a 2D motion flow $\vec{u}_p = (u_p, v_p)$ on the projector film and the corresponding flows on the screen and the viewer's retina are denoted by $\vec{u}_s$ and $\vec{u}_v$, respectively. From equations (4.2) and (4.3), we have

$$\vec{u}_p = \frac{f_p \vec{u}_s}{D_p} \tag{4.4}$$

$$\vec{u}_v = \frac{f_v \vec{u}_s}{D_v} = \frac{f_v D_p \vec{u}_p}{D_v f_p} = \frac{\vec{u}_p}{k}. \tag{4.5}$$

Equation (4.5) suggests that the flow $\vec{u}_v$ perceived on the retina is scaled by a factor k compared with the corresponding flow $\vec{u}_p$ on the projector film. The flow $\vec{u}_p$ is given by

$$\begin{aligned} u_p &= \tfrac{W}{Z}\left(x_p - f_c \tfrac{U}{W}\right) + \alpha \tfrac{x_p y_p}{f_c} - \beta\left(\tfrac{x_p^2}{f_c} + f_c\right) + \gamma y \\ v_p &= \tfrac{W}{Z}\left(y_p - f_c \tfrac{V}{W}\right) - \beta \tfrac{x_p y_p}{f_c} + \alpha\left(\tfrac{y_p^2}{f_c} + f_c\right) - \gamma y \end{aligned} \tag{4.6}$$

where the 3-D motion parameters (U, V, W) and (α, β, γ) represent the motion experienced by the director's camera, and f_c is the focal length of the directors camera. Expanding the horizontal component of $\vec{u}_v$ in equation (4.5) and bringing

in equation (4.6), we obtain:

$$\begin{aligned} u_v &= \frac{W}{Z}\left(x_v - \frac{f_c U}{kW}\right) + \alpha\frac{x_v y_v}{\frac{f_c}{k}} \\ &\quad -\beta\left(\frac{x_v^2}{\frac{f_c}{k}} + \frac{f_c}{k}\right) + \gamma y_v \end{aligned} \tag{4.7}$$

Similar expression can be written for the vertical component of the flow v_v. From equation (4.7), we see that the flow field experienced by the viewer indirectly through the screen is one that arises from the same external motion and depths experienced by the director's camera, i.e. $U, V, W, \alpha, \beta, \gamma,\ Z$, but with a modified focal length $f_v' = \frac{f_c}{k} = f_v\frac{D_p f_c}{D_v f_p}$. Thus only when the viewer is seated at the CVP $\left(D_v = \frac{D_p f_p}{f_c}\right)$, the motion field $\vec{u}_v$ is undistorted (i.e. the same as that experienced by a viewer making the 3-D motion himself/herself).

Clearly, if the viewer is able to revise the estimate of its intrinsic parameter from f_v to f_v', he/she is then no worse off than the case of having to solve the SFM problem when experiencing an undistorted 2-D motion flow. Even the viewer is not able to estimate the new focal length, we shall show later that the effect of this focal length error is benign, as far as scene structure recovery is concerned.

The above analysis can now be extended to the case where the viewer's and the projector's optical axes are parallel but not coincident (4.1(b)). If the viewer is located at a position $(t_x, t_y, D_c - D_v)$ away from the CVP, as illustrated in Figure 4.2, there will be a shift in the principal point (o_x, o_y). Since $\frac{o_x}{t_x} = -\frac{f_v}{D_v}$ and $\frac{o_y}{t_y} = -\frac{f_v}{D_v}$, we have $(o_x, o_y) = \left(-t_x\frac{f_v}{D_v}, -t_y\frac{f_v}{D_v}\right)$. Thus the optical flow can be

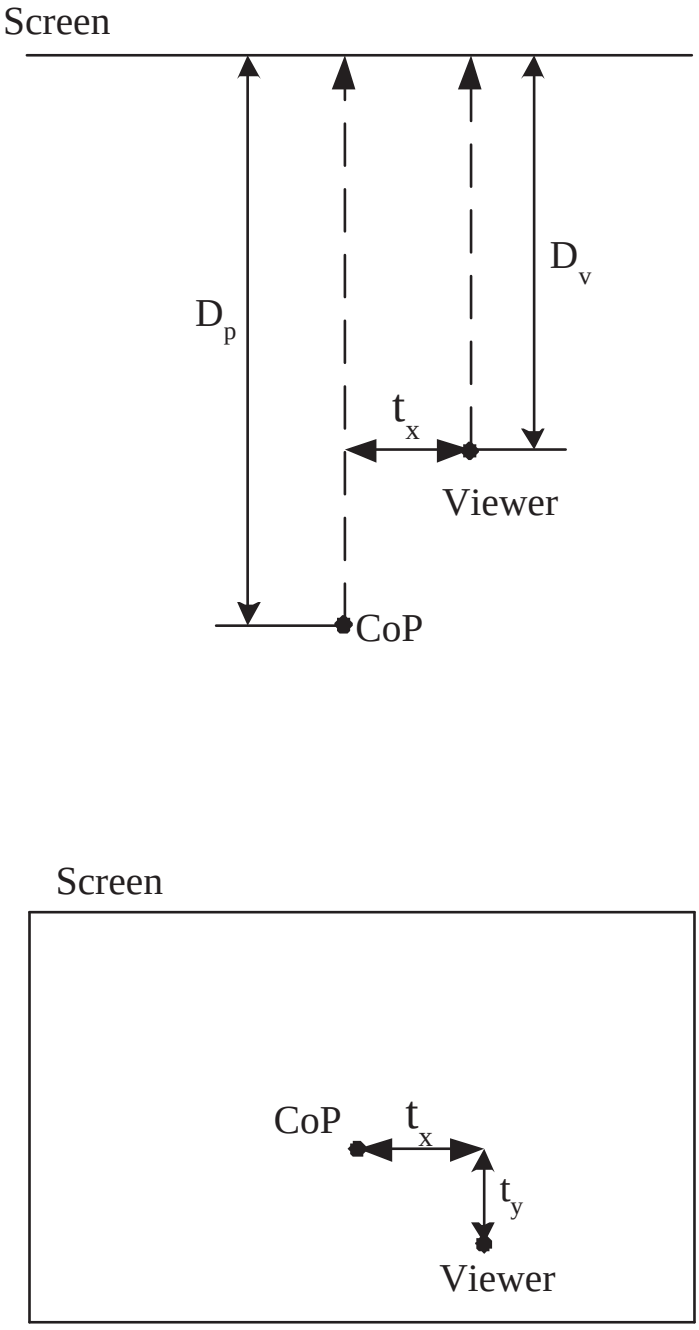


Figure 4.2: The configuration where the viewer's and projector's optical axes are parallel but not coincident.

written in the following form:

$$\begin{aligned} u_v = \frac{W}{Z}\left(x_v + t_x\frac{f_v}{D_v}\right) - f'_v\frac{U}{Z} + \frac{\left(x_v+t_x\frac{f_v}{D_v}\right)\left(y_v+t_y\frac{f_v}{D_v}\right)}{f'_v}\alpha \\ -f'_v\left(1+\frac{\left(x_v+t_x\frac{f_v}{D_v}\right)^2}{f'^2_v}\right)\beta + \gamma\left(y_0 + t_y\frac{f_v}{D_v}\right) \end{aligned} \tag{4.8}$$

$$\begin{aligned} v_v = \frac{W}{Z}\left(y_v + t_y\frac{f_v}{D_v}\right) - f'_v\frac{V}{Z} - \frac{\left(x_v+t_x\frac{f_v}{D_v}\right)\left(y_v+t_y\frac{f_v}{D_v}\right)}{f'_v}\beta \\ +f'_v\left(1+\frac{\left(y_v+t_y\frac{f_v}{D_v}\right)^2}{f'^2_v}\right)\alpha - \gamma\left(x_v + t_x\frac{f_v}{D_v}\right) \end{aligned} \tag{4.9}$$

This is similar to the optical flow that would be obtained if the principal point of the viewer's optical system is not $(0,0)$ as we have assumed so far, but given by $\left(t_x\frac{f_v}{D_v}, t_y\frac{f_v}{D_v}\right)$. In sum, for the simple scenario where the viewer's and the projector's optical axes are parallel, the motion estimation problem is no more complex than an uncalibrated SFM problem, where in particular the focal length f'_v might be different over time due to the director using different lenses, but over most of the time, f'_v would not be dynamically varying unless the director is using the zooming shot. The difference with the usual uncalibrated SFM problem is that the principal point offset $\left(t_x\frac{f_v}{D_v}, t_y\frac{f_v}{D_v}\right)$ can be very much larger (especially $t_x\frac{f_v}{D_v}$) than one usually encounters in computer vision problem.

In general, if the intrinsic parameters are not calibrated, the visual system cannot recover the Euclidean geometry of the scene from motion cues, but only its projective or affine geometry [29]. The question remains whether a person "calibrates" his or her visual system. Indeed, the anatomy of the eye varies from the fovea to the periphery and such parameters also change over time. Computationally, the

estimation of intrinsic parameters from the motion cues is possible, even for varying focal length and principal point [46]. However, it is numerically ill-conditioned and always requires higher-order constraints. It seems that such higher-order mechanisms cannot explain how the brain may estimate eye intrinsic parameters [105]. It could be that the visual system only needs to obtain a very rough estimation of the intrinsic parameters, and knowing these rough estimates, it is sufficient to obtain certain aspect of depth information [12]. If this were indeed the case, then what the cinema viewer experiences is just a more severe version of the situation with bigger errors in the estimates for the intrinsic parameters. We shall see in Section 4.4 how these errors in the intrinsic parameters (as well as the extrinsic parameters) will affect spatial perception. In particular, we will compare the changes caused by the large principal point offset $\left(t_x \frac{f_v}{D_v}, t_y \frac{f_v}{D_v}\right)$ in the characteristics of the depth distortion. More importantly, we show that despite these errors, certain key qualitative properties of the recovered depth remain unchanged.

4.3.2 Optical axes of viewer and projector not parallel

In general, movie viewers could watch the screen with a slant angle if he/she does not sit right on the optical axis of the projector (see Figure 4.3). To relate how the dynamic cues are transformed as a result of the slant, we relate the view of a viewer seated at CVP and that of a viewer seated at a general position via a

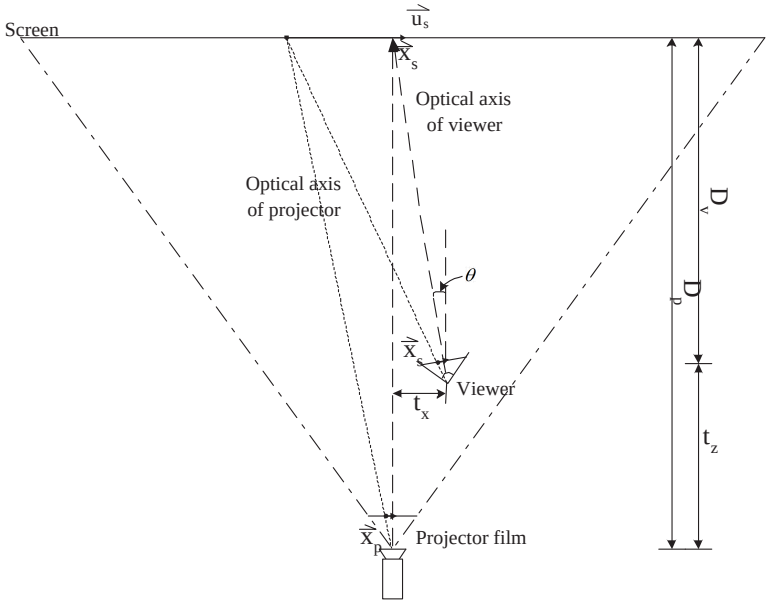


Figure 4.3: A general configuration, with a slant ϕ in the viewer's optical axis around the vertical axis.

homography (a 3×3 linear transformation) induced by the screen plane [43]. Here we again assume that the axis of the projector is perpendicular to the screen. In other words, no keystone distortion [78] is present in the cinema. Thus the unit vector normal to the screen plane is given by $\mathbf{N} = (0, 0, -1)^T$. Then the coordinates of a feature point $\mathbf{x_v}$ of a viewer seated at CVP and $\mathbf{x'_v}$ of a viewer seated at the general position can be related by a simple homography $\mathbf{H}$:

$$\mathbf{x'}_\mathbf{v} = \mathbf{H}\mathbf{x}_\mathbf{v} \tag{4.10}$$

where the homography is given by [43]

$$\mathbf{H} = \mathbf{K'_v}\left(\mathbf{R} + \frac{1}{D_c}\mathbf{t}\mathbf{N}^T\right)\mathbf{K_v^{-1}}. \tag{4.11}$$

In the above equation, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ denote the rigid transformation between the viewer and the CVP. We have assumed that human eye is modeled by a pinhole camera; thus $\mathbf{K'}_{\mathrm{v}}$ and $\mathbf{K}_{\mathrm{v}}$ are the intrinsic parameter matrices characterizing the eyes of the viewer seated at the CVP and the general position respectively, with the form of $\mathbf{K'}_{\mathrm{v}}$ and $\mathbf{K}_{\mathrm{v}}$ given by that of $\mathbf{K}$ in equation (2.3). We assume that s_θ, o_{xv}, and o_{yv} for the eyes at both positions to be 0. We also assume that the focal lengths f_v for the eyes at both positions to be identical, because given the typical distance of the screen, both focal lengths will correspond to the eyes at the most relaxed state. Thus:

$$\mathbf{K}_{\mathrm{v}} = \begin{bmatrix} f_v & 0 & 0 \\ 0 & f_v & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.12}$$

We assume that the viewer is gazing at a region near the central part of the screen, and we first derive the simple case where the viewer is at the same vertical level as the projector. There is an angle of ϕ between the viewer's optical axis and the vertical axis but there is no rotation around the X-axis. Conceptually, one can also reduce any rotations around both X- and Y-axes to a single rotation about the Y-axis by a suitable in-plane rotation of the $X-Y$ coordinate axes. Referring to Figure 4.3, the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ can thus be written

as:

$$\mathbf{R} = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix} \tag{4.13}$$

$$\mathbf{t} = \begin{bmatrix} t_x & 0 & t_z \end{bmatrix}^T \tag{4.14}$$

The homography $\mathbf{H}$ is then readily obtained as:

$$\mathbf{H} = \mathbf{K_v} \begin{bmatrix} \cos\phi & 0 & -\sin\phi - \frac{t_x}{D_c} \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi - \frac{t_z}{D_c} \end{bmatrix} \mathbf{K_v}^{-1} \tag{4.15}$$

Using the projection model of equation (2.3), the whole projection process can then be written as:

$$\mathbf{x'}_\mathbf{v} = \mathbf{H}\mathbf{x_v} = \mathbf{H}\mathbf{K}_\mathrm{v} \left[\mathbf{I}|\mathbf{0}\right] \mathbf{X} \tag{4.16}$$

where $\mathbf{X}$ is the 3-D point that gives rise to $\mathbf{x_v}$ ($\mathbf{x_v}$, $\mathbf{x'_v}$, $\mathbf{X}$ all expressed in homogeneous coordinates).

Equation (4.16) shows that there is a new "intrinsic parameter matrix" $\mathbf{H'} = \mathbf{H}\mathbf{K}_\mathrm{v}$ underlying the image formation process of a cinema viewer seated at a general position. Unfortunately the"intrinsic parameter matrix" induced by the homography

is not an upper-triangular matrix like that of a typical intrinsic parameter matrix:

$$\mathbf{H}' = \mathbf{H}\mathbf{K}_{\mathrm{v}} = \begin{bmatrix} \cos\phi f_v & 0 & -f_v \sin\phi - \frac{t_x}{D_c} f_v \\ 0 & f_v & 0 \\ \sin\phi & 0 & \cos\phi - \frac{t_z}{D_c} \end{bmatrix} \tag{4.17}$$

If ϕ is sufficiently small, we can simplify $\mathbf{H}'$ such that the effect of ϕ can be regarded as a perturbation to $\mathbf{K}_{\mathrm{v}}$. Firstly, the lowest right entry can be approximated by

$$\cos\phi - \frac{t_z}{D_c} \approx \frac{D_c - t_z}{D_c} = \frac{D_v}{D_c}. \tag{4.18}$$

Since $\mathbf{H}'$ is up to an arbitrary scale factor that is inherent in homogeneous representation, we can scale the whole matrix such that the lowest right entry is unity.

$$\mathbf{H}' = \mathbf{H}\mathbf{K}_{\mathbf{v}} = \begin{bmatrix} \cos\theta f_v \frac{D_c}{D_v} & 0 & -\frac{t_x}{D_v} f_v \\ 0 & f_v \frac{D_c}{D_v} & 0 \\ \sin\theta \frac{D_c}{D_v} & 0 & 1 \end{bmatrix} \tag{4.19}$$

Clearly, if ϕ is small enough such that $\cos\phi \approx 1$ and $\sin\phi \approx 0$, then the intrinsic parameter matrix reduces to that of section 4.3.1. However, in the general case, the matrix does not have the typical form for intrinsic parameter matrix in view of the non-zero lowest left entry. Then, (x'_v, y'_v) can be written as

$$x'_v = \frac{\cos\theta f_v \frac{D_c}{D_v} X - \frac{t_x}{D_v} f_v Z}{Z + \sin\theta \frac{D_c}{D_v} X} \tag{4.20}$$

$$y'_v = \frac{f_v Y}{Z + \sin\theta \frac{D_c}{D_v} X} \tag{4.21}$$

The projection at the general position can thus be regarded as one with not only changes in focal length and principal point offset, but now these changes also vary

in magnitudes from the fovea to the periphery (the denominators in equations (4.20) and (4.21) change as X increases from fovea to periphery). In other words, the projection rays do not intersect at one point, with the result that we have a non-central projection system [92, 91]. How would this impact on the viewer seated at this general position? The human visual system is itself prey to non-ideal effects like spherical aberration, coma and other asymmetries expected from a biological system. For instance, the optical surfaces may lack rotational symmetry and their nominal centres of curvature may not lie on a common axis; such meridional changes in radius of curvature lead to ocular astigmatism [9].

Though these aberrations occur in the human eyes, its visual effect is minimal. For instance, the astigmatic image falls on the peripheral retina which has relatively poor resolving power compared to the retina at the macula. Thus peripheral spatial vision performance seems little affected, though the effect of these off-axis errors on spatial and temporal sampling in the periphery is not yet completely determined, with some recent works being [33, 2, 39].

The question here is whether there is a need for the visual system to recalibrate a system (to whatever extent) with such more severe aberrations introduced. The key depends on the type of space recoverable or indeed being recovered by the human visual system, under both everyday SFM and under cinema viewing condition. We are now coming to the central question of depth distortion under both situations

in the next section.

4.4 Depth distortion arising from erroneous estimation of 3-D motion and intrinsic parameters

4.4.1 Iso-distortion framework

The iso-distortion framework was first introduced by Cheong et al. [10]. The iso-distortion framework seeks to understand the geometric laws under which the recovered scene is distorted due to some errors in the estimated camera parameters. This is motivated by the fact that it is unlikely for a human visual system to recover the exact motion parameters and hence it is important to understand how the perceived space is distorted by such errors in the motion estimates.

Referring to equation (4.1), we note that if there are errors in the estimates of the extrinsic parameters, these errors will in turn cause errors in the estimation of the scaled depth. To simplify the discussion, we assume there is no error in the optical flow, since we are primarily concerned with how errors in the motion parameters affect depth reconstruction. Plugging the various motion estimates and

the expression for the optical flow (equations (2.5) and (2.6) into equation (4.1) , we obtain the distorted depth $\hat{Z}$ as follows:

$$\hat{Z} = Z\left(\frac{(x-\hat{x}_0, y-\hat{y}_0)\cdot\mathbf{n}}{(x-x_0, y-y_0)\cdot\mathbf{n} + Z\,(u_{rot_e}, v_{rot_e})\cdot\mathbf{n}}\right), \tag{4.22}$$

Equation (4.22) shows that errors in the motion estimates distort the recovered relative depth by a factor D, given by the terms in the bracket, which among other terms, contains the term $\mathbf{n}$. As mentioned in the discussion following equation (4.1), the value of $\mathbf{n}$ depends on the scheme we use to recover depth. In our work, we choose to recover depth along the estimated epipolar direction, i.e. $\mathbf{n} = \frac{(x-\hat{x}_0, y-\hat{y}_0)^T}{\sqrt{(x-\hat{x}_0)^2+(y-\hat{y}_0)^2}}$. Such a choice is reasonable because the estimated epipolar direction contains the strongest translational flow and hence is the most reliable direction to recover Z. Hence the distortion factor D becomes:

$$D = \frac{(x-\hat{x}_0)^2 + (y-\hat{y}_0)^2}{(x-x_0, y-y_0)\cdot(x-\hat{x}_0, y-\hat{y}_0) + Z\,(u_{rot_e}, v_{rot_e})\cdot(x-\hat{x}_0, y-\hat{y}_0)}. \tag{4.23}$$

The complexity of equation (4.23) can be intuitively grasped with a graphical approach in its first analysis. For specific values of the parameters $x_0,\ y_0,\ \hat{x}_0,\ \hat{y}_0,\ \alpha_e,\ \beta_e,\ \gamma_e$ and for any fixed distortion factor D, equation (4.23) describes a surface $g\,(x, y, Z) = 0$ in the xyZ-space. The entire ensemble of such surfaces, each for a different value of D, describes the distortion action of the motion errors on any points in the 3-D space. Normally, under general motion, a complicated distortion characteristic may arise. Readers are referred to [10, 12] for a full description of the geometry of the distortion.

Algebraically, it was shown from [10] that given such motion errors, the transformation from the physical to the perceived space belongs to the family of Cremona transformations, whereby the homogeneous coordinates of a point in the perceived space $\left[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}, \hat{\mathcal{W}}\right]$ is related to the actual point $[\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}]$ by:

$$\left[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}, \hat{\mathcal{W}}\right] = [\phi_1, \phi_2, \phi_3, \phi_4]$$

where the quantities ϕ_i are homogeneous polynomials in $[\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}]$. Such transformation is bijective almost everywhere except on the set of what is known as fundamental elements where the correspondence between the two spaces becomes one-to-many[48]. The complex nature of this transformation makes it clear that in general it is very difficult to recover metric depth accurately. What is less clear is the feasibility of recovering some of the less metrical depth representations under specific motions. For instance, the ordinal representation of depth constitutes one such reduced representation of depth where only depth order is available. Cheong and Xiang [12] showed that, though small amount of motion errors can have significant impact on depth recovery in the general case, there exist generic motions that allow robust recovery of such partial depth information. In particular, lateral motion is better than forward motion in terms of yielding ordinal depth information and other aspects of depth recovery, in spite of the fact that the ambiguity between the camera rotation and translation is more severe in this case. On the other hand, forward motion leads to conditions more conducive for 3-D motion

estimation compared to the case of lateral motion, but it is not necessarily good for depth recovery. This dichotomy between forward and lateral motion means that it is important for a biological system to choose a motion intelligently so as to accomplish tasks robustly.

In the case of uncalibrated motion with fixed intrinsic parameters and reasonably small principal point offset, the distortion factor D becomes[12]:

- for lateral motion ($W = 0$):

$$D = \frac{\hat{f}\hat{U}}{fU + \left(\beta f - \hat{\beta}\hat{f}\right) Z} \tag{4.24}$$

- for forward motion ($U = V = 0$):

$$D = \frac{x^2 + y^2}{\left((x - o_{xe})\, x + (y - o_{ye})\, y\right) + \left(-\left(\beta f - \hat{\beta}\hat{f}\right) x + \left(\alpha f - \hat{\alpha}\hat{f}\right) y\right) Z}. \tag{4.25}$$

It was shown in [12] that lateral motion is better than forward motion in terms of yielding ordinal depth information, in spite of the fact that the ambiguity between the camera rotation and translation is more severe in this case. On the other hand, forward motion leads to conditions more conducive for 3-D motion estimation compared to the case of lateral motion, but it is not necessarily good for depth recovery. The aforementioned properties regarding the dichotomy in depth and motion recovery are not affected, in spite of possible errors in the intrinsic

parameters. However, if the intrinsic parameters are allowed to vary dynamically (equations for D under such case not shown here; see [12] for a fuller account), then even ordinal depth information might not be recoverable under lateral motion.

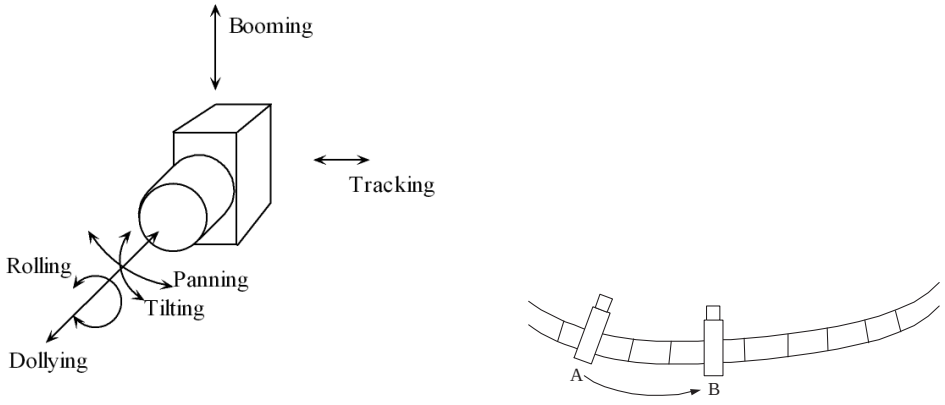
The upshot of characterizing depth distortion behaviour under these generic types of forward and lateral motions are the following two aspects: (1) It shows that the reliability of a reconstructed scene has quite a different behaviour from that of the motion estimates. For instance, if the motion contains dominant lateral translation, it might be very difficult to lift the ambiguity between translation and rotation. However, in spite of such motion ambiguity, certain aspect of depth information seems recoverable with robustness. Indeed, in the biological world, lateral motions are often executed to judge distance and relative ordering. On the other hand, psychophysical experiments [104] reported that under pure forward translation, human subjects were unable to recover structure unless favorable conditions such as large field of view exist. Thus it seems that not all motions are equal in terms of robust depth recovery and that there also exists certain dichotomy between forward and lateral translation as far as motion and depth recovery are concerned. (2) Understanding the depth recovered under these two very different motion types gives us an epistemological idea about the geometry of the perceived space under general motions, in the sense that the behaviour of depth reconstruction at these two opposite poles of translation spectrum delimits the type of general depth distortion behaviour somewhere in between the two poles. Clearly, in the absence of

other depth cues, or without using additional scene knowledge, Euclidean or even affine depth recovery may not be possible in general.

4.4.2 Depth distortion in cinema

We now apply the iso-distortion framework to look at the SFM problem under the cinema viewing configuration. Like previous iso-distortion analyses, we restrict ourselves to scenes where only the camera is moving or we assume that in scenes where there are independently moving objects, these objects have been properly segmented. We focus on the situation depicted in Figure 4.1(b) which has been shown to be equivalent to an uncalibrated SFM problem for the viewer, with mostly fixed but possibly unknown focal length and potentially very large principal point offset.

One might ask to what extent the notion of generic motion employed in the previous analyses is valid or relevant in the cinema context. In cinematography, camera motions are not arbitrary, but are dictated by the need to communicate meanings and by the mechanics of film-making. For instance, a panning shot is often used to establish the scenes of a new shot and to track an object or person. A dolly shot (translation in depth; see Fig 4.4(a)) is used to move in closer to a subject or to effect a first-person viewpoint shot as the protagonist moves forward. Shots with more complex combinations of motions are possible, for instance, translation and



(a) (b)

Figure 4.4: Camera operations: (a) basic terminologies for translational and rotational operations, (b) typical camera operation on rail.

rotation are often coupled together in tracking shots using the setup illustrated in Figure 4.4(b). Nevertheless, it is reasonable to say that in terms of translation, the shots either exhibit primarily forward/backward translation or primarily lateral translation. Thus, consistent with the assumption made in the previous paper [12], we can hypothesize that the viewer is at least aware what generic type of motion is being executed by the camera. That is, the motion estimates are such that

- for lateral motion, $\hat{W} = W = 0$; and
- for forward motion, $\hat{U} = U = \hat{V} = V = 0$.

We ignore zooming motion and its possible confusion with forward translation. Even though zoom lenses are prevalent nowadays, the experience of zooming motion is not a natural phenomenon to our eyes. Excessive zooming in or out may

irritate the viewer and hence, zooming is not commonly used, except in some cases where special effects are required [67, 58]. For instance, in the film Vertigo (1958), Hitchcock makes Scotty's illness visible and intelligible through the simultaneous combination of a forward zoom and a dollying out (backward translation), this "combination of approach and retreat whose complex confusions of perspective briefly induce all the sensations of nausea in the spectator" [82]. This rare use of zooming is fortunate as it is difficult to separate the flow field induced by a zooming-in from the flow field simultaneously created by a forward translation. It also justifies our decision to ignore such motion in our analysis. Next, we also assume that the contribution of γ_e is very small. Camera operations in cinematography usually minimize rotation about the optical axis (rolling) so as to avoid causing excessive discomfort to viewers. Lastly, in our first presentation of the distortion characteristics, we make an assumption that will allow us to better grasp the major geometrical features of the depth distortion: within a limited field of view, second order rotational terms in the image co-ordinates are small relative to the linear and constant terms. This is the case when the visual system focuses its attention on the fovea region under normal viewing condition. Even if this assumption is removed, given their typical magnitudes, these terms do not qualitatively affect the nature of the depth distortion. However in the cinema viewing configuration, we will reinstate those second order terms caused by the principal point offset $(t_x \frac{f_v}{D_v}, t_y \frac{f_v}{D_v})$ as the latter is large and no longer negligible.

4.4.3 Lateral motion

If we assume that viewer is aware of the type of generic motion being made, then under lateral motion all $\mathbf{n}$ will be in the same direction given by $\mathbf{n} = -\frac{\left(\hat{U},\hat{V}\right)^T}{\sqrt{\hat{U}^2+\hat{V}^2}}$, for the epipolar reconstruction scheme of recovering depth. For notational convenience, we can rotate the X- and Y-axes without loss of generality so that $\mathbf{n}$ becomes $(1,0)^T$ or $\left(\hat{U},\hat{V}\right)$ lies in the direction $(1,0)$ (though (U,V) need not lie in that direction). Thus the optical flow caused by lateral motion can be written as

$$u_v = -f'_v\frac{U}{Z} + \frac{\left(x_v - o'_x\right)\left(y_v - o'_y\right)}{f'_v}\alpha - f'_v\left(1 + \frac{\left(x_v - o'_x\right)^2}{f'^2_v}\right)\beta + \gamma\left(y_v - o'_y\right) \quad (4.26)$$

$$v_v = -f'_v\frac{V}{Z} - \frac{\left(x_v - o'_x\right)\left(y_v - o'_y\right)}{f'_v}\beta + f'_v\left(1 + \frac{\left(y_v - o'_y\right)^2}{f'^2_v}\right)\alpha - \gamma\left(x_v - o'_x\right) \quad (4.27)$$

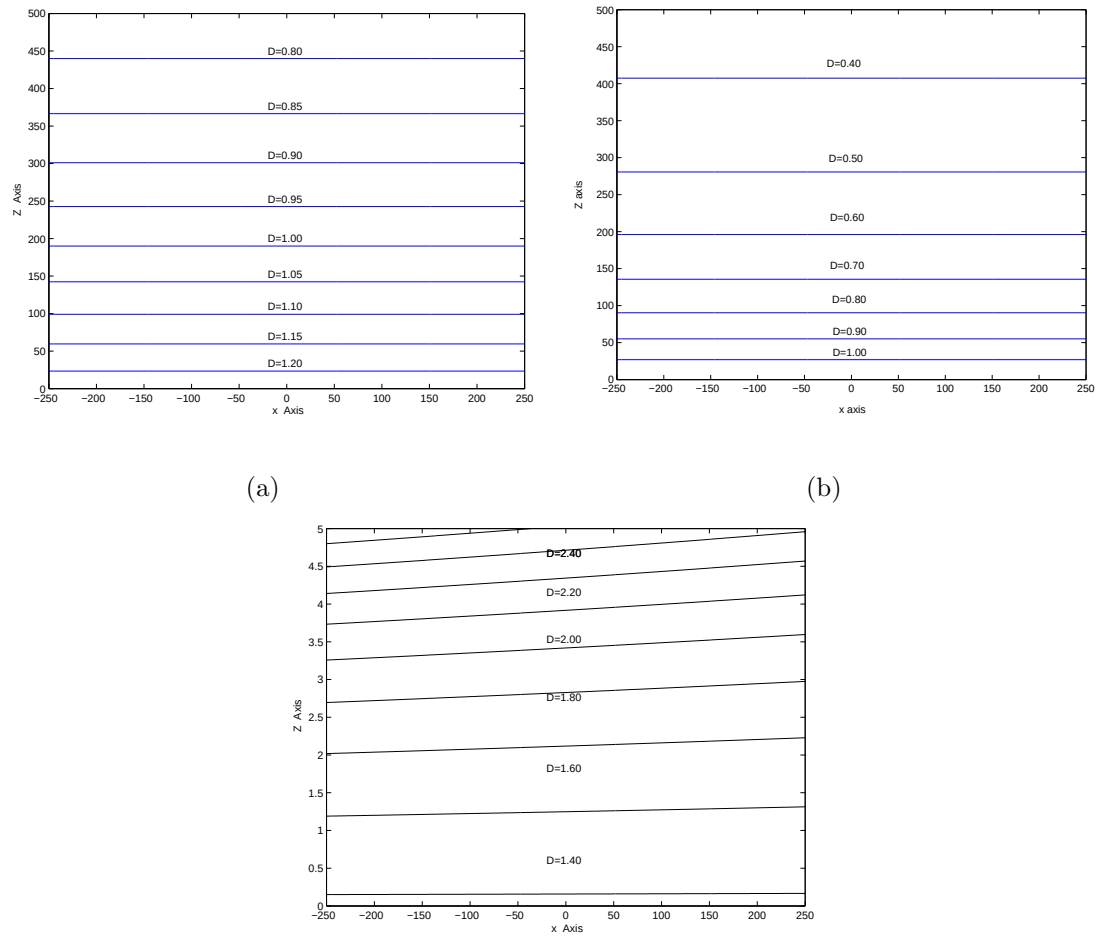
where

$$\left(o'_x, o'_y\right) = \left(-t_x\frac{f_v}{D_v}, -t_y\frac{f_v}{D_v}\right).$$

Plugging in the value of $\mathbf{n}$, the optical flow given by equations (4.26) and (4.27), and the estimated quantities $\hat{x}_0,\ \hat{y}_0,\ \hat{\alpha},\ \hat{\beta},\ \hat{\gamma},\ \hat{o}'_x,\ \hat{o}'_y,\ \hat{f}'_v$ into equation (4.24) we obtain the distortion factor D:

- for the case of uncalibrated SFM under normal viewing condition (second order terms due to rotation and principal point ignored):

$$D = \frac{\hat{f}'_v\hat{U}}{f'_vU + \left(\beta f'_v - \hat{\beta}\hat{f}'_v\right)Z} \quad (4.28)$$



(a) (b)

(c)

Figure 4.5: Families of iso-distortion contours for lateral motion obtained by intersecting the iso-distortion surfaces with the xZ-plane. $FoV = 53^o$, $f = f'_v = 309.0$, $U = V = 0.81$, $\beta = -0.002$, $\alpha = 0.002$,. (a) Viewer at CVP with errors only in the 3-D motion estimates, $\hat{U} = 1.0$, $\hat{\beta} = -0.001$ (b) Viewer with optical axis parallel to and coincident with the projector's optical axis $\hat{U} = 1.0$, $\hat{\beta} = -0.001$ $\hat{f}'_v = 303.0$ (c) Viewer in a general viewing position. $\hat{U} = 1.0$, $\hat{V} = 1.0$, $\hat{\beta} = -0.001$, $\hat{\alpha} = 0.001, \hat{f}'_v = 303.0, o'_x = o'_y = 10000$.

- for the case of cinema viewing configuration (second order terms due to principal point dominant):

$$D = \frac{\hat{f}'_v \hat{U}}{f'_v U + \left(\left(\beta f'_v - \hat{\beta}\hat{f}'_v\right) + O^2\left(x_v, y_v\right)\right) Z} \tag{4.29}$$

where

$$\begin{aligned} O^2\left(x_v, y_v\right) &= \beta\frac{\left(x_v - o'_x\right)^2}{f'_v} - \alpha\frac{\left(x_v - o'_x\right)\left(y_v - o'_y\right)}{f'_v} \\ &\quad -\hat{\beta}\frac{\left(x_v - \hat{o}'_x\right)^2}{\hat{f}'_v} + \hat{\alpha}\frac{\left(x_v - \hat{o}'_x\right)^2\left(y_v - \hat{o}'_y\right)}{\hat{f}'_v} \end{aligned} \tag{4.30}$$

The distortion factor expressed in equation (4.28) for normal viewing condition has the form $\frac{1}{a+bZ}$, where $a = \frac{f'_v U}{\hat{f}'_v \hat{U}}$ and $b = \frac{\beta f'_v - \hat{\beta}\hat{f}'_v}{\hat{f}'_v \hat{U}}$ are constants for all the scene points. It has the property that the distortion preserves the depth order of any two recovered depths $\hat{Z}_1$ and $\hat{Z}_2$ under certain conditions that are likely to hold (see [12] for details). For instance, if $Z_1 > Z_2$, it can be readily shown that, given either of the following conditions, depending on the sign of a:

- $\left(a + bZ_1\right)\left(a + bZ_2\right) > 0$ if $a > 0$, or
- $\left(a + bZ_1\right)\left(a + bZ_2\right) < 0$ if $a < 0$

the transformation $\hat{Z} = DZ$ preserves the depth order of the two points, that is, $\hat{Z}_1 > \hat{Z}_2$. Since $a = \frac{f'_v U}{\hat{f}'_v \hat{U}}$, the condition $a > 0$ means that $f'_v U$ and $\hat{f}'_v \hat{U}$ have the

same sign. This condition can easily be met by human visual system; thus we can just focus on the first condition. The requirement $(a + bZ_1)(a + bZ_2) > 0$ simply means that the two estimated depths should have the same sign. This condition can be easily assured by checking the sign of $\hat{Z}_1$ and $\hat{Z}_2$. If they are of the same sign, the depth order of $\hat{Z}_1$ and $\hat{Z}_2$ is correct; otherwise, reverse the depth order. Furthermore, if the errors in the motion estimates are small enough, then this perceived ordinal depth space converges to a metric space.

Now consider equation (4.29). It is of a similar form $\frac{1}{a+bZ}$, but with b given by the non-constant expression:

$$b = \frac{\left(\beta f'_v - \hat{\beta}\hat{f}'_v\right) + O^2(x_v, y_v)}{\hat{f}'_v \hat{U}}$$

where $O^2(x_v, y_v)$ is given by equation (4.30). Clearly, ordinal depth is not preserved since the value of b depends on (x_v, y_v). However, if the offset terms o'_x, o'_y, $\hat{o}'_x$, and $\hat{o}'_y$ in $O^2(x_v, y_v)$ dominate (x_v, y_v), then b remains largely the same over a local region, and to the extent that b is constant, the ordinality of depths recovered within this local region is likely to be preserved. See Fig 4.5 for the values of D in the $x - Z$ plane under various viewing positions. We make the following observations

- The sign of b decides whether the perceived space is compressed or expanded (compared Figures 4.5(a), 4.5(b) with 4.5(c)), with the depth order preserved irrespective of the sign of b.

- There is no qualitative difference between the distortion in Figure 4.5(a) and Figure 4.5(b), despite the addition of second order rotational terms (which results in the bending of the contour) and the error in the focal length. This echoes the result of our paper [12] that calibration is not the determining factor in the quality of the perceived space.

- With the large principal point offset error found in the cinema viewing condition, the bending is made more pronounced by the second order terms arising from this offset, which is further aggravated by the shift in the origin. This results in difficulty in deciding depth orders across large visual angle, which seems to be consistent with out experience of sitting in an extreme off-center position.

4.4.4 Forward motion

For the case of forward motion, adopting the same "epipolar reconstruction" scheme, $\mathbf{n}$ can be expressed as $\frac{(x,y)^T}{\sqrt{x^2+y^2}}$. The distortion factor D can then be expressed as:

- for the case of normal uncalibrated SFM (with all second order terms ignored)

$$D = \frac{\left(x_v - \hat{o}'_x\right)^2 + \left(y_v - \hat{o}'_y\right)^2}{\left(x_v - o'_x\right)\left(x_v - \hat{o}'_x\right) + \left(y_v - o'_y\right)\left(y_v - \hat{o}'_y\right) + O'Z} \tag{4.31}$$

  where

$$O' = -\left(\beta f'_v - \hat{\beta}\hat{f}'_v\right)\left(x - \hat{o}'_x\right) + \left(\alpha f'_v - \hat{\alpha}\hat{f}'_v\right)\left(y - \hat{o}'_y\right) \tag{4.32}$$

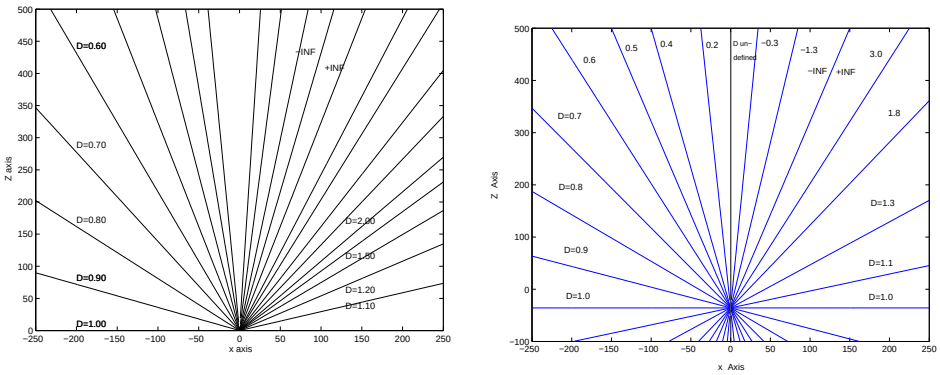
- for the case of cinema viewing configuration (second order terms due to principal point dominant)

$$D = \frac{(x-\hat{o}'_x)^2 + \left(y-\hat{o}'_y\right)^2}{(x_v-o'_x)(x_v-\hat{o}'_x) + \left(y_v-o'_y\right)\left(y_v-\hat{o}'_y\right) + O''Z} \tag{4.33}$$

where

$$\begin{aligned}
O'' &= -\widehat{\beta}_f (x-\hat{o}'_x) + \widehat{\alpha}_f \left(y-\hat{o}'_y\right) \\
\widehat{\beta}_f &= \beta f'_v - \hat{\beta}\hat{f}'_v + O_x^2(x_v, y_v) \\
\widehat{\alpha}_f &= \alpha f'_v - \hat{\alpha}\hat{f}'_v + O_y^2(x_v, y_v) \\
O_x^2(x_v, y_v) &= -\beta\frac{(x_v-o'_x)^2}{f'_v} + \alpha\frac{(x_v-o'_x)\left(y_v-o'_y\right)}{f'_v} \\
&\quad +\hat{\beta}\frac{(x_v-\hat{o}'_x)^2}{\hat{f}'_v} - \hat{\alpha}\frac{(x_v-\hat{o}'_x)\left(y_v-\hat{o}'_y\right)}{\hat{f}'_v} \\
O_y^2(x_v, y_v) &= -\alpha\frac{\left(y_v-o'_y\right)^2}{f'_v} + \beta\frac{(x_v-o'_x)\left(y_v-o'_y\right)}{f'_v} \\
&\quad +\hat{\alpha}\frac{\left(y_v-\hat{o}'_y\right)^2}{\hat{f}'_v} - \hat{\beta}\frac{(x_v-\hat{o}'_x)\left(y_v-\hat{o}'_y\right)}{\hat{f}'_v}
\end{aligned}$$

From both equations (4.31) and (4.33), we see that D cannot be expressed in the form of $\frac{1}{a+bZ}$ with constant a and b. Indeed, for a particular value of D, the corresponding iso-distortion surface is a cone. It has also been shown [11] that all D surfaces in the 3-D space intersect on a common line. As can be seen the distortion factor varies rapidly in a small neighborhood (Fig 4.6(a) and 4.6(b)) around the forward direction, and thus depth reconstruction is much more difficult than that in the case of lateral motion. While the presence of the second order terms may change the shape of the iso-distortion contours towards the periphery,



(a) (b)

Figure 4.6: Families of iso-distortion contours for forward motion. (a) Viewer seated at CVP, $f_v = 309.0,\ \beta_e = 0.001,\ \alpha_e = 0.001$ (b) Viewer seated on the optical axis of the projector with $D_v < D_c,\ f'_v = 309.0,\ \hat{f}'_v = 303.0,\ \beta = -0.002,\ \hat{\beta} = -0.001,\ \alpha = 0.002,\ \hat{\alpha} = 0.001$. INF stands for infinity.

the key properties discussed above regarding depth distortion are still true. In particular, ordinal depths are no longer recoverable. On the contrary, it has been shown [14] that forward motion leads to conditions favorable for motion recovery.

In sum the discussion so far in this section has shown that while the multiple projection processes in a cinema viewing configuration (with optical axis of viewer and projector being parallel) may vary the iso-distortion equations, they do not alter the essential properties of depth distortion for both lateral and forward motion. In other words, since in the first place it is difficult even under normal viewing condition to obtain exact motion estimates from motion cues, the key properties of the perceived depths are already laid down, with changes in the intrinsic parameters

(brought about by the cinema viewing condition) contributing only to quantitative but not qualitative change.

4.5 Discussion

Various psychophysical experiments have showed that we cannot recover the Euclidean space from two views even in our everyday activities. This is manifest in various psychophysical phenomena such as apparent frontal parallel plane (AFPP), apparent distance bisection (ADB), and foreshortening of visual space at increasing distance under stereo vision [73] (note that human stereopsis is mathematically equivalent to a lateral monocular translation along the inter-ocular distance, followed by an eye rotation equal to the convergence angle). AFPP has also been reported for the case of motion [13]. This inability to recover the veridical space is also mirrored by the computational difficulties encountered in depth reconstruction algorithms from motion and stereo cues. In particular, the resultant distortion in the recovered depth is modeled by the distribution of the iso-distortion surfaces presented in this thesis. For instance, Figures 4.5(a) and 4.5(b) explain the compression of stereoscopic space noted by various researchers [35, 61, 107]. The surprising thing is that we function remarkably well in everyday life and this seeming paradox parallels that happening in the cinema.

The results of our work showed the link between everyday SFM and that occurring in the cinema. In particular, viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems. What are the implications of these results? Is there a need to calibrate these changes in the intrinsic parameters? It is an open question whether the human visual system does this. There is no need to calibrate if, in the first place, we are not even able to estimate extrinsic motion parameters accurately under everyday SFM condition. Such errors in the motion parameters render Euclidean space recovery impossible and in fact already determine all the important properties of space distortion. Changes or errors in the intrinsic parameters introduce further changes in the perceived shape but the qualitative nature of distortion remains the same.

Clearly, without a comprehensive psychophysical investigation, we cannot say conclusively about the nature of space representation used by the human visual system. However the epistemological considerations (what can and what cannot be recovered) raised by this computational inquiry do constrain the likely forms of space recovered from motion cues over two views. It seems that that recovery of metrical depth information is in general very difficult; indeed, even recovery of partial depth such as ordinal depth information might not be possible under all situations. Having now at our disposal the various computational results regarding depth perception under cinema viewing configuration and uncalibrated SFM in general, we are testing these predictions with psychophysical experiments so as to confirm or

refine our views about the role of calibration in human vision.

Let us explore the cinematographic implications even we do suffer from depth distortion arising from motion cues (not considering the role played by other cues). Firstly, from our discussion in Section 4.4 about the nature of generic motions, it means that the establishment shots favored by directors to introduce scenes will yield reliable ordinal depth information, because of the lateral motions employed in these shots. This is true irrespective of whether there is calibration of the intrinsic parameters or not, and as long as the seat position is not too far off to the side. Such qualitative appreciation of the scene depth might be sufficient to render cinematic communication between the director and the audience possible.

On the other hand, shots with primarily forward motion present conditions favorable for motion recovery but not for depth recovery, regardless of whether the intrinsic parameters are calibrated or not [14]. Such shots are mainly used in closing in towards a subject or to effect a first-person view as he or she navigates through some environment. In the latter scenarios, the ability to recover the direction of motion well is obviously important for the appreciation of the meaning of the shot. Aspects of structural information might also be important for the viewer to "inhabit" the space of the protagonist, although its recovery from motion cues might not be feasible. Which particular structural aspect needs to be recovered is task-dependent; for instance, the ability to estimate the time-to-collision (TTC) is important for shots depicting chases, say, through tight corridors. Fortunately,

such information can be recovered directly from the optical flow, without going through the step of 3D motion recovery [57, 70, 93].

Finally, it must be added that even though the distortion may seem severe for two-frame SFM, the viewing conditions experienced by human being are typically not so impoverished in depth cues, be it in everyday life or in the cinema. For instance, merely extend the SFM problem to multiple views and the recovered structure has to obey the constraint of rigidity. Other cues such as static perspective cue play an important role too. The work by Stevens and Brookes [87], Sparrow and Stine [86] or Cornilleau-Pèrés et al [20] have shown that static cues can dominate stereopsis or motion cues for the perception of plane orientation. Cutting [21] showed that the nonrigidity predicted by motion cue for a viewer not seated at the CVP is not perceived and one explanation is that the static cues overrule the motion cue. Indeed, static cues might also be used to recover H, the homography that relates the view of a person seated at the CVP to that seated at a general position. For instance, the orthogonality assumption among the detected vanishing directions enable partial self-calibration of the principal point from just a single view.

Chapter 5

Conclusions and Future Work

5.1 The behavior of SFM with erroneous intrinsic parameters

After intensive research in the past two decades, the geometric and computational aspect of SFM seems well studied. However, the state of the art is that a practical SFM algorithm that can handle general visual tasks in the real world is still unavailable. One of the contributing reasons is that the first step of SFM involves solving an ill-conditioned problem. The computation of feature correspondence or optical flow is under-constrained in nature, thus additional assumptions such as depth smoothness are needed. Therefore, the input for the second and third steps

is inevitably contaminated by errors which in turn lead to a distorted reconstruction. Consequently, the focus of SFM research has been shifted among others to the robustness and sensitivity issues in recent years. In [75], Oliensis proposes a new critique of SFM research. He argues that more comprehensive theoretical as well as phenomenological analyses of algorithm behavior should be carried out under all sort of typical scenarios. Such analyses are important not only for understanding algorithms' properties, but also for conducting good experiments and for developing the best algorithms. Our work is toward this direction. The analysis about the motion estimation with erroneous focal length is based on [110]. In particular, we are concerned with the limitation of SFM algorithms in the face of errors in the estimation of the focal length. Instead of dealing with specific algorithms each using different optimization techniques, we study one class of algorithms based on the weighted differential epipolar constraint. The error surfaces under a wide range of motion-scene configurations are studied and plotted, from which several results are drawn.

In our work we have developed expressions describing the error behavior of egomotion estimation when the focal length is calibrated with error. The key results are independent of both the egomotion estimation as well as the calibration algorithms. We show the bas-relief valley will be rotated according to the error in the focal length, in a way that is dependent on the motion-scene configuration. One important suggestion is that, provided that one knows the rough range of the true

focal length, setting a larger-than-true focal length helps to estimate the direction of translation better though possibly with larger biases.

The results also show that the effect of erroneous focal length on the FOE estimate is not the same over different translation and rotation directions. The structure of the scene (depth) affects the shifting of the FOE estimate as well.

For the case of varying calibration parameters (f dynamically changing), additional analyses are in order. The results established in [12]—that zoom field crucially influence properties of depth reconstruction —raise the possibility that the results might be quite different.

5.2 How movie viewers perceive scene structure from dynamic cues

Our work offered an analytic account of several properties of the perceived visual space when viewing the cinema from a location other than the CoP. In section 4.3, we prove that the dynamic perception of pictures viewed from this location with optical axis parallel to the projector's axis can be treated as one where the viewer experiences a change in the intrinsic parameters. Such changes remain within the framework of uncalibrated SFM proposed for machine vision, and thus the viewer

can use an algorithm similar to the various self-calibration algorithms proposed in the computational vision community, if such algorithm exists at all in the human brain. If the viewing axis and the projector axis are not parallel, then such viewing configuration not only changes the intrinsic parameters, but the amount of changes themselves are a function of the eye's eccentricity, a situation not dissimilar to the complex geometry of a foveated eye.

Even if the viewer is not able to calibrate these intrinsic parameters, we show that the situation is not as serious as it seems. We investigate the properties of the depth recovery and find that the ability of depth recovery is not jeopardized under the cinema configuration. In other words, the estimation errors of the intrinsic parameters will not change the essential properties of depth recovery. Lateral motion still leads to robust ordinal depth recovery, whereas for forward motion, the chief factor contributing to severe distortion in depth recovery is the difficulty in estimating the extrinsic parameters well enough.

5.3 Future Work

The problem of recovering the structure of a 3-D scene from a sequence of images obtained from the relative motion between the scene and the observer is an important area of research in computer. As we mentioned earlier, only a few subproblems

of the SFM problems have been addressed in this thesis. The theoretical framework and methodology adopted in our work can be extended to tackle the other aspects of the SFM problem.

In the first part of thesis, we presents detailed geometric analysis, along with simulations, of the errors computed for a large class of SFM estimation algorithms. The analysis focuses on the errors caused by estimating the intrinsic characteristics of the camera (specifically the focal length of the camera) inaccurately. The results suggest that error on the side of larger focal lengths might result in more accurate estimates of directions of motions. However, smaller focal lengths might result in more stable estimations of motion. This may inspire new algorithms of ego-motion and camera calibration algorithms. What so ever, we have also show certain combination of the direction translation and rotation may help to reduce the effect of erroneous intrinsic parameters on the ego-motion estimation.

The second part of our research relates the problem of perceiving motion on the screen of a cinema to the errors analyzed in the first part. We show that this viewpoint discrepancy may be modeled as an error in estimating the intrinsic parameters of the human system and suggests that depth recovery from motion cues are not jeopardized much. Our geometrically motivated approach for understanding the calibrated motion ambiguities can be readily extended to deal with other viewing condition. For example Head-mounted displays (HMDs) and 3DTV, where the the modified focal length focal length is more close to that of our visual system.

Appendix A

Decomposition of Homography Matrix

Among the various catalogues of explanations to the cinema viewing paradox, the compensation hypotheses have lots of proponents. These hypotheses claim that invariance is the consequence of re–interpreting the retinal image by recovering the position of the CoP from either the information in the picture. Thus the CoP recovery is a major problem for these hypotheses. Most of the algorithms agree that the CoP is recovered from the locations of vanishing points in the light field. It has been proved that the locations of three orthogonal vanishing points are sufficient to recover the vanishing point [8, 54]. Alternatively, two orthogonal vanishing points plus the assumption that the CoP lies ont the surface normal from the center of

the picture can be used to recover CoP [84]. In this section we propose another possible algorithm that many be used to recovered CoP.

Considering the case discussed in Section 4.3.2, if the viewer is aware of the translation and rotation of himself/herself with respect to the CoP, clearly the position of CoP can be recovered easily. Assume the viewer can somehow recover the homography matrix $\mathbf{H}$ either from the pictorial compensation or from the global surface compensation, then whether $\mathbf{H}$ can be decomposed into its motion and components parameters, namely $\left\{\mathbf{R}, \frac{\mathbf{t}}{D_p}, \mathbf{N}\right\}$ is the major problem for the above hypothesis.

We now prove that given a homography matrix $\mathbf{H} = \left(\mathbf{R} + \frac{1}{D_p}\mathbf{t}\mathbf{N}^T\right)$, there are at most two physically possible solutions for a decomposition into its components $\left\{\mathbf{R}, \frac{\mathbf{t}}{D_p}, \mathbf{N}\right\}$.

First note that $\mathbf{H}$ preserves the length of any vector orthogonal to $\mathbf{N}$. Also, if we know the plane spanned by the vectors that are orthogonal to $\mathbf{N}$, we then know $\mathbf{N}$ itself. We first recover the vector $\mathbf{N}$ based on this knowledge.

The symmetric matrix $\mathbf{H}^T\mathbf{H}$ has three eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq 0$ with $\sigma_2 = 0$. Since $\mathbf{H}^T\mathbf{H}$ is symmetric, it can be diagonalized by a 3×3 orthogonal matrix $\mathbf{V}$ such that:

$$\mathbf{H}^T\mathbf{H} = \mathbf{V}\sum\mathbf{V}^T \tag{A.1}$$

where $\sum = diag\left\{\sigma_1^2, \sigma_2^2, \sigma_3^2\right\}$. We denote the three column vectors of $\mathbf{V}$ as $[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$;then

we have:

$$\mathbf{H}^T\mathbf{H}\mathbf{v}_1 = \sigma_1^2\mathbf{v}_1 \tag{A.2}$$

$$\mathbf{H}^T\mathbf{H}\mathbf{v}_2 = \sigma_2^2\mathbf{v}_2 \tag{A.3}$$

$$\mathbf{H}^T\mathbf{H}\mathbf{v}_3 = \sigma_3^2\mathbf{v}_3 \tag{A.4}$$

Since $\mathbf{v_2}$ is orthogonal to both $\mathbf{N}$ and $\mathbf{t}$, and its length is preserved under the map $\mathbf{H}$. Also, it is easy to check that the lengths of two other unit-length vectors defined as

$$\mathbf{u}_1 = \frac{\sqrt{1-\sigma_3^2}\mathbf{v}_1 + \sqrt{\sigma_1^2-1}\mathbf{v}_3}{\sqrt{\sigma_1^2-\sigma_3^2}},\ \mathbf{u}_2 = \frac{\sqrt{1-\sigma_3^2}\mathbf{v}_1 - \sqrt{\sigma_1^2-1}\mathbf{v}_3}{\sqrt{\sigma_1^2-\sigma_3^2}} \tag{A.5}$$

is also preserved under the map $\mathbf{H}$. Furthermore, it is easy to verify that $\mathbf{H}$ preserves the length of any vectors inside each of the two subspaces

$$S_1 = span\left\{\mathbf{v}_2, \mathbf{u}_1\right\},\ S_2 = span\left\{\mathbf{v}_2, \mathbf{u}_2\right\} \tag{A.6}$$

Since $\mathbf{v_2}$ is orthogonal to $\mathbf{u_1}$ and $\mathbf{u_2}$, $[\mathbf{v}_2]_\times \mathbf{u}_1$ is a unit normal vector to S_1, and $[\mathbf{v}_2]_\times \mathbf{u}_2$ is a unit normal vector to S_2. Define the matrices:

$$\begin{aligned} \mathbf{U}_1 &= \left[\mathbf{v}_2, \mathbf{u}_1, [\mathbf{v}_2]_\times \mathbf{u}_1\right],\ \mathbf{W}_1 = \left[\mathbf{H}\mathbf{v}_2, \mathbf{H}\mathbf{u}_1, [\mathbf{H}\mathbf{v}_2]_\times \mathbf{H}\mathbf{u}_1\right] \\ \mathbf{U}_2 &= \left[\mathbf{v}_2, \mathbf{u}_2, [\mathbf{v}_2]_\times \mathbf{u}_2\right],\ \mathbf{W}_2 = \left[\mathbf{H}\mathbf{v}_2, \mathbf{H}\mathbf{u}_2, [\mathbf{H}\mathbf{v}_2]_\times \mathbf{H}\mathbf{u}_2\right] \end{aligned} \tag{A.7}$$

We then have

$$\mathbf{R}\mathbf{U}_1 = \mathbf{W}_1,\ \mathbf{R}\mathbf{U}_2 = \mathbf{W}_2. \tag{A.8}$$

This suggests that each subspace S_1 or S_2 may give rise to a solution to the decomposition where $\mathbf{R}$ is given by $\mathbf{W_1}\mathbf{u_1^T}$ or $\mathbf{W_1}\mathbf{u_2^T}$. By taking into account

the sign ambiguity in the term $\frac{1}{D_p}\mathbf{t}\mathbf{N}^T$, we obtain four solutions for decomposing $\mathbf{H} = \left(\mathbf{R} + \frac{1}{D_p}\mathbf{t}\mathbf{N}^T\right)$ to $\left\{\mathbf{R}, \frac{\mathbf{t}}{D_p}, \mathbf{N}\right\}$:

1. $\mathbf{R}_1 = \mathbf{W}_1\mathbf{U}_1^T,\ \mathbf{N}_1 = [\mathbf{v}_2]_\times \mathbf{u}_1,\ \frac{1}{D_p}\mathbf{t}_1 = (\mathbf{H} - \mathbf{R}_1)\,\mathbf{N}_1$

2. $\mathbf{R}_2 = \mathbf{W}_2\mathbf{U}_2^T,\ \mathbf{N}_2 = [\mathbf{v}_2]_\times \mathbf{u}_2,\ \frac{1}{D_p}\mathbf{t}_2 = (\mathbf{H} - \mathbf{R}_2)\,\mathbf{N}_2$

3. $\mathbf{R}_3 = \mathbf{R}_1,\ \mathbf{N}_3 = -\mathbf{N}_1,\ \frac{1}{D_p}\mathbf{t}_3 = -\frac{1}{D_p}\mathbf{t}_1$

4. $\mathbf{R}_4 = \mathbf{R}_2,\ \mathbf{N}_4 = -\mathbf{N}_2,\ \frac{1}{D_p}\mathbf{t}_4 = -\frac{1}{D_p}\mathbf{t}_2$

In order to reduce the number of physically possible solutions, we may impose the positive depth constraint. For example, if solution 1 is the true one; this constraint will then eliminate solutions 3 as being physically impossible. Similarly, one of the solutions 2 or 4 will be eliminated. Thus, we end up with at most two solutions.

Bibliography

[1] G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field. *PAMI*, 11(5):477–489, May 1989.

[2] P. Artal, A. M. Derrington, and E. Colombo. Refraction, aliasing, and the absence of motion reversals in peripheral vision. *Vision Research*, 35(7):939, 1995.

[3] M. S. Banks, H. F. Rose, D. Vishwanath, and A.R. Girshick. Where should you sit to watch a movie? In *SPIE: Human Vision and Electronic Imaging*, 2005.

[4] P. A. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure from motion. In *ECCV94*, pages B:85–96, 1994.

[5] S. Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *ICCV98*, pages 790–796, 1998.

[6] M.J. Brooks, W. Chojnacki, and L. Baumela. Determining the egomotion of an uncalibrated camera from instantaneous optical flow. *JOSA-A*, 14(10):2670–2677, October 1997.

[7] A. R. Bruss and B. K. P. Horn. Passive navigation. *CVGIP*, 21(1):3–20, January 1983.

[8] B. Caprile and V. Torre. Using vanishing points for camera calibration. *IJCV*, 4(2):127–140, March 1990.

[9] W. N. Charman. *Visual optics and instrumentation*, chapter Optics of the human eye. Macmillan Press, 1991.

[10] L. F. Cheong, C. Fermuller, and Y. Aloimonos. Effects of errors in the viewing geometry on shape estimation. *CVIU*, 71(3):356–372, September 1998.

[11] L. F. Cheong and C. H. Peh. Depth distortion under calibration uncertainty. *CVIU*, 93(3):221–244, March 2004.

[12] L. F. Cheong and T. Xiang. Characterizing depth distortion under different generic motions. *IJCV*, 44(3):199–217, September 2001.

[13] L. F. Cheong, T. Xiang, V. Cornilleau-Pérèz, and Tai L. C. Not all motions are equivalent in terms of depth recovery. In John X. Liu, editor, *Computer Vision and Robotics*, 2005.

[14] L. F. Cheong and X. Xiang. Error characteristics of SFM with unknown focal length. In *ACCV*, 2006.

[15] A. Chiuso, R. Brockett, and S. Soatto. Optimal structure from motion: Local ambiguities and global estimates. *IJCV*, 39(3):195–228, September 2000.

[16] M. Christian. *Film Language: A Semiotics of the Cinema. Trans. Michael Taylor*. New York: Oxford University Press, 1974.

[17] K. Cornelis, M. Pollefeys, M. Vergauwen, and L. Van Gool. Augmented reality using uncalibrated video sequences. *Lecture Notes in Computer Science*, 2018:144–160, 2001.

[18] V. Cornilleau-Pèrés and J. Droulez. Visual perception of surface curvature: Psychophysics of curvature detection induced by motion parallax. *Perception and Psychophysics*, 46(4):351–364, 1989.

[19] V. Cornilleau-pérès and J. Droulez. The visual perception of 3D shape from self-motion and object-motion. *Vision Research*, 34:2331–2336, 1994.

[20] V. Cornilleau-Pèrés, M. Wexler, J. Droulez, E. Marin, C. Miège, and B. Bourdoncle. Visual perception of planar orientation: dominance of static depth cues over motion cues. *Vision Research*, 42:1403-12, 2002.

[21] J. E. Cutting. Rigidity of cinema seen from the front row, side aisle. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):323–334, 1987.

[22] W. J. M. Damme, F. H. Oosterhoff, and W. A. van de Grind. Discrimination of 3-D shape and 3-D curvature from motion in active vision. *Perception and Psychophysics*, 55:340–349, 1994.

[23] W. J. M. Damme and W. A. van de Grind. Active vision and the identification of 3D shape. *Vision Research*, 11:1581–1587, 1993.

[24] K. Daniilidis and M. E. Spetsakis. Understanding noise sensitivity in structure from motion. In *VisNav93*, 1993.

[25] F. Domini, C. Caudek, and Richmann S. Distortions of depth-order relations and parallelism in structure from motion. *Perception and Psychophysics*, 60:1164–1174, 1998.

[26] J. Droulez and V. Cornilleau-Pérès. Visual perception of surface curvature, the spin varaition and its physiological implications. *Biological Cybernetics*, 62(3):211–224, 1990.

[27] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *ECCV92*, pages 563–578, 1992.

[28] O. D. Faugeras. *Three-Dimensional Computer Vision*. MIT press, 1993.

[29] O. D. Faugeras. Stratification of 3D vision: Projective, affine and metric representations. *JOSA-A*, 12(3):465–484, March 1995.

[30] O. D. Faugeras, Q. T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *ECCV92*, pages 321–334, 1992.

[31] O. D. Faugeras and G. Toscani. The calibration problem for stereo. In *CVPR86*, pages 15–20, 1986.

[32] C. Fermuller and Y. Aloimonos. Observability of 3D motion. *IJCV*, 37(1):43–63, June 2000.

[33] S. J. Galvin, D. R. Williams, and N. J. Coletta. The spatial grain of motion perception in human peripheral vision. *Vision Research*, 36(15):2283–2296, 1996.

[34] J. J. Gibson. *The perception of the visual world*. Boston: Houghton Mifflin, 1950.

[35] W. C. Gogel. *Foundations of perceptual*, chapter The analysis of perceived space, pages 113–182. Amsterdam: North-Holland, 1993.

[36] J. I. González, J. C. Gámez, C. G. Artal, and A. M. N. Cabrera. Stability study of camera calibration methods. *CI Workshop en Agentes Físicos, WAF, Spain*, 2005.

[37] M. A. Goodale and D. A. Westwood. An evolving view of duplex vision: separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology*, 14(2):203–211, 2004.

[38] E. Grossmann and J. Santos-Victor. Uncertainty analysis of 3-D reconstruction from uncalibrated views. *IVC*, 18(9):685–696, June 2000.

[39] A. Guirao and P. Artal. Off-axis monochromatic aberrations estimated from double pass measurements in the human eye. *Vision Research*, 39(2):207–217, 1999.

[40] R. I. Hartley. Projective reconstruction and invariants from multiple images. *PAMI*, 16(10):1036–1041, October 1994.

[41] R. I. Hartley. In defense of the eight-point algorithm. *PAMI*, 19(6):580–593, June 1997.

[42] R. I. Hartley and C. Silpa-Anan. Reconstruction from two views using approximate calibration. In *ACCV*, pages 338–343, 2002.

[43] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[44] D. J. Heeger and A. D. Jepson. Linear subspace methods for recovering translation direction. In *RBCV-TR*, 1992.

[45] D. J. Heeger and A. D. Jepson. Subspace methods for recovering rigid motion i: Algorithms and implementation. *IJCV*, 7(2):95–117, January 1992.

[46] A. Heyden and K. Åström. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *CVPR*, pages 438–443, 1997.

[47] W. C. Hoffman. The lie algebra of visual perception. *Journal of Mathematical Psychology*, 3:65–98, 1966.

[48] H. P. Hudson. *Cremona Transformations in Plane and Space. Cambridge.* Cambridge: Cambridge University Press, Cambridge, 1927.

[49] Y. Sugaya K. Kanatani, A. Nakatsuji. Stabilizing the focal length computation for 3d reconstruction from two uncalibrated views. *IJCV*, 66(2):109–122, 2006.

[50] F. Kahl and B. Triggs. Critical motions in euclidean structure from motion. In *CVPR99*, pages II: 366–372, 1999.

[51] K.I. Kanatani. 3-D interpretation of optical-flow by renormalization. *IJCV*, 11(3):267–282, December 1993.

[52] J. J. Koenderink and A. J. van Doorn. Relief: Pictorial and otherwise. *Image and Vision Computing*, 13(5):321–334, 1995.

[53] E. Kruppa. Zur ermittlung eines objektes aus zwei perspektiven mit innerer orientierung. *Sitz.-Ber. Akad. Wiss., Math. Naturw., Kl. Abt. IIa*, 122:1939–1948, 1913.

[54] M. Kubovy. *The Psychology of Perspective and Renaissance Art.* Cambridre University Press, New York, 1986.

[55] J. Z. C. Lai. On the sensitivity of camera calibration. *IVC*, 11(10):656–664, December 1993.

[56] J. M. Lavest, M. Viala, and M. Dhome. Do we really need an accurate calibration pattern to achieve a reliable camera calibration? In *ECCV98*, page I: 158, 1998.

[57] D. N. Lee. The optic flow field: The foundation of vision. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290(1038):169–178, Jun 1980.

[58] B. Leeuw. *Digital Cinematography.* AP Professional, 1997.

[59] D. Liebowitz and A. Zisserman. Resolving ambiguities in auto-calibration. *Royal*, A-356:1193–1211, 1998.

[60] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

[61] J. M. Loomis and J. W. Philbeck. Is the anisotropy of perceived 3-D shape invariant across scale? *Perception and Psychophysics*, 61:397–402, 1999.

[62] B. D. Lucas. *Generalized image matching by the method of differences.* PhD thesis, Carnegie-Mellon University, 1984.

[63] R. K. Luneburg. *Mathematical analysis of binocular vision*. Princeton, NJ: Princeton University Press, 1947.

[64] Q. T. Luong, R. Deriche, O. D. Faugeras, and T. Papadopoulo. On determining the fundamental matrix: Analysis of different methods and experimental results. In *INRIA*, 1993.

[65] J. Košecká M. Zucchelli. Motion bias and structure distortion induced by intrinsic calibration errors. *IVC*, 2007.

[66] Y. Ma, J. Kosecka, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *IJCV*, 44(3):219–249, September 2001.

[67] M. Martha. *Producing Videos: A Complete Guide*. AFTRS, Sydney, 1997.

[68] S. J. Maybank and O. D. Faugeras. A theory of self-calibration of a moving camera. *IJCV*, 8(2):123–151, August 1992.

[69] T. S. Meese and M. G. Harris. Computation of surface slant from optic flow: orthogonal components of speed gradient can be combined. *Vision Research*, 37(17):2369–2379, 1997.

[70] R. C. Nelson and J. Aloimonos. Obstacle avoidance using flow field divergence. *PAMI*, 11(10):1102–110, Oct 1989.

[71] J. Neumann, C. Fermuller, and Y. Aloimonos. A hierarchy of cameras for 3D photography. *CVIU*, 96(3):274–293, December 2004.

[72] J. F. Norman and J. S. Lappin. The detection of surface curvatures defined by optical motion. *Perception and Psychophysics*, 51(4):386–396, 1992.

[73] K. N. Ogle. *Researches in Binocular Vision.* New York: Hafner, 1964.

[74] J. Oliensis. A multi-frame structure-from-motion algorithm under perspective projection. *IJCV*, 34(2-3):163–192, August 1999.

[75] J. Oliensis. A critique of structure-from-motion algorithms. *CVIU*, 80(2):172–214, November 2000.

[76] J. Oliensis. A new structure-from-motion ambiguity. *PAMI*, 22(7):685–700, July 2000.

[77] D. N. Perkins. Compensating for distortion in viewing pictures obliquely. *Perception and Psychophysics*, 14:13–18, 1973.

[78] R. Petrozzo and S. W. Singer. Cinema projection distortion. In *The 141st SMPTE Technical Conference and Exhibition*, 1999.

[79] M. H. Pirenne. *Optics, painting, and photography.* Cambridge, England: Cambride University Press, 1970.

[80] R. Pless. Using many cameras as one. In *CVPR*, pages II: 587–593, 2003.

[81] M. Pollefeys, R. Koch, and L. J. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *ICCV98*, pages 90–95, 1998.

[82] D. Raymond. *The Strange Case of Alfred Hitchcock*. Cambridge, Mass.: M.I.T. Press, 1978.

[83] B. J. Rogers and M. E. Graham. Anisotropies in the perception of three-dimensional surfaces. *Science*, 221, 1983.

[84] H. A. Sedgwik. The effects of viewpoint on the virtual space of pictures. *Pictorial Communication in Virtual and Real Environments*, 1991.

[85] S. W. Shih, Y. P. Hung, and W. S. Lin. Accuracy assessment on camera calibration method not considering lens distortion. In *CVPR92*, pages 755–757, 1992.

[86] J. E. Sparrow and W. M. Stine. The perceived rigidity of rotating eight-vertex geometric forms; extracting nonrigid structure from rigid motion. *Vision Research*, 38:541–556, 1998.

[87] K. A. Stevens and A. Brookes. Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research*, 28:371–386, 1998.

[88] P. F. Sturm. Critical motion sequences for monocular self-calibration and uncalibrated euclidean reconstruction. In *CVPR97*, pages 1100–1105, 1997.

[89] P. F. Sturm. Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length. In *BMVC99*, pages 63–72, 1999.

[90] P. F. Sturm. On focal length calibration from two views. In *CVPR01*, pages II:145–150, 2001.

[91] P. F. Sturm. Multi-view geometry for general camera models. In *CVPR05*, pages I: 206–212, 2005.

[92] P. F. Sturm and S. Ramalingam. A generic concept for camera calibration. In *ECCV04*, pages Vol II: 1–13, 2004.

[93] M. Subbarao. *Interpretation of Visual Motion: A Computational Study*. Pitman Publishing Limited, 1988.

[94] T. Svoboda and P. Sturm. What can be done with a badly calibrated camera in ego-motion estimation? In *TR*, 1996.

[95] R. Szeliski and S. B. Kang. Shape ambiguities in structure-from-motion. *PAMI*, 19(5):506–512, May 1997.

[96] J. S. Tittle, J. T. Todd, V. J. Perotti, and Norman J. F. Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance*, 21(33):663–678, 1995.

[97] J. T. Todd and P. Bressan. The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception and Psychophysics*, 48:419–430, 1990.

[98] J. T. Todd and V. J. Perotti. The visual perception of surface orientation from optical flow. *Perception and Psychophysics*, 61(8):1577–1589, 1999.

[99] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9(2):137–154, November 1992.

[100] B. Triggs. Autocalibration from planar scenes. In *ECCV*, page I: 89, 1998.

[101] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.

[102] R. Y. Tsai. An efficient and accurate camera calibration technique for 3-D machine vision. In *CVPR86*, pages 364–374, 1986.

[103] R. Y. Tsai and T. S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. In *PRIP82*, pages 112–118, 1982.

[104] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.

[105] T. Viéville, J. Droulez, C. H. Peh, and A. Negri. How do we perceive the eye intrinsic parameters? *RR 4030: INRIA Technical Report*, 2000.

[106] T. Viéville, O. D. Faugeras, and Q. T. Luong. Motion of points and lines in the uncalibrated case. *IJCV*, 17(1):7–41, January 1996.

[107] M. Wagner. The metric of visual space. *Perception and Psychophysics*, 38(6):483–495, 1985.

[108] A. M. Waxman, B. Kamgar-Parsi, and M. Subbarao. Closed form solutions to image flow equations for 3D structure and motion. *IJCV*, 1:239–258, 1987.

[109] J. Weng, N. Ahuja, and T. S. Huang. Optimal motion and structure estimation. *PAMI*, 15(9):864–884, September 1993.

[110] T. Xiang and L. F. Cheong. Understanding the behavior of SFM algorithms: A geometric approach. *IJCV*, 51(2):111–137, February 2003.

[111] G. S. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field. *PAMI*, 14(10):995–1013, October 1992.

[112] Z. Y. Zhang. On the optimization criteria used in two-view motion analysis. *PAMI*, 20(7):717–729, July 1998.

[113] Z. Y. Zhang. Understanding the relationship between the optimization criteria in two-view motion analysis. In *ICCV*, pages 772–777, 1998.

[114] Z. Y. Zhang, Q. T. Luong, and O. D. Faugeras. Motion of an uncalibrated stereo rig: Self-calibration and metric reconstruction. *RA*, 12(1):103–113, February 1996.

[115] X. Zhuang and R.M. Haralick. Rigid body motion and optical flow image. In *the 1st International Conferrence on Artificial Intelligence Application*, pages 366–375, 1984.

[116] X. Zhuang, T. S. Huang, N. Ahuja, and R.M. Haralick. A simplified linear optical flow-motion algorithm. *CVGIP*, 42(3):334–344, June 1988.

Publication List

L. F. Cheong and X. Xiang How do we Perceive Depths from Motion Cues in the Movies: A Computational Account, accepted by Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2007

L. F. Cheong and X. Xiang Error characteristics of SFM with unknown focal length. In ACCV, 2006.

L. F. Cheong and X. Xiang How do Movie Viewers perceive scene Structure from Dynamic cues. In CVPR, 2006.

X. Xiang, C. D. Cheng, C. C. Ko and B. M. Chen, S.J. Lu, "API for virtual laboratory instrument using Java3D," the 3rd International Conference on Control Theory and Applications, Pretoria, South Africa, December 2001.

X. Xiang, C. C. Dong, C. C. Ko, and B. M. Chen, "Development of Web-Based 3D Oscilloscope Experimentation" SingAREN Symposium 02, March 2002.