LINEAR QUASI-PARALLAX S*f*M FOR VARIOUS CLASSES OF BIOLOGICAL EYES

Chuanxin Hu

NATIONAL UNIVERSITY OF SINGAPORE

2011

LINEAR QUASI-PARALLAX S*f*M FOR VARIOUS CLASSES OF BIOLOGICAL EYES

Chuanxin Hu

(B.Eng. (Electronic Engineering), SJTU)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF

ELECTRICAL & COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2011

To my parents Mr. Wei HU and Mrs. Xiufeng XU

To my sister Ms. Mei XU

Acknowledgments

Acknowledgments

PhD study is a long and lonely journey: we venture into unknown research areas and search around hoping to find a research topic that is promising but not yet discovered by fellow researchers. It is like mining gold in a desert; usually it takes many fruitless attempts before we spot one potential gold mine. Then we start to dig with high hopes, soon realizing that all the previous hard work merely marked the beginning of a long journey. There are more roadblocks and challenges on the way that I often felt frustrated, disheartened and inclined to give up halfway. Thus, with all sincerity, I shall say it is impossible for me to finish my PhD study, without the encouragement, inspiration and support from my supervisor, family, friends and labmates.

I feel deeply indebted to my supervisor Dr. Cheong Loong-fah, who offered me great advice and guidance throughout my PhD study. His teaching and mentoring turned computer vision into a fun, interesting and stimulating subject to study. Moreover, his enthusiasm and deep care for the nature and biology inspired me a lot.

I'm also grateful to my fellow students in the VIP lab for their comraderie and support through difficult times. And to our lab technician Francis for making our lab a very pleasant environment to study and for helping me in many ways.

My gratitude also goes to my friends at Microsoft for their friendly nudges reminding me to finish my thesis and for their good faith in me even in the craziest times.

Lastly and most importantly, I want to thank my parents who not only gave me life and raised me, but also taught me valuable life lessons and brought out the best in me. Always being there for me, my little sister is another strong force behind me. I dedicated my thesis to them for their unconditional love showered on me.

Contents

Summary

A large class of visual systems in the biological world often has multiple eyes in simultaneous motion and yet has little or no overlap in the visual fields between the eyes. These systems include the lateral eyes found in many vertebrates and the compound eyes in insects. Instead of computing feature correspondences between the eyes, which might not even be possible due to the lack of overlap in the visual fields, we exploit the organizational possibility offered by the eye topography. In particular, we leverage on the pair of visual rays that are parallel to each other but opposite in direction, and compute what we call the quasiparallax for translation recovery. Besides resulting in parsimonious visual processing, the quasi-parallax term also enhances the information pick-up for the translation, as it is almost rotation-free. The rotation is subsequently recovered from a pencil of visual rays using the individual epipolar constraints of each camera. As a result of using these different and appropriate aspects of visual rays for motion recovery, our method is numerically more effective in disambiguating the translation and rotation. In comparison to the gold standard solution obtained by the bundle adjustment (BA) technique, our method has a better Fisher information matrix for a lateral eye pair, as well as a superior experimental performance under the case of narrow field of view. For other eye configurations, the two methods achieve comparable performances, with our linear method slightly edging the nonlinear BA method when there exists imperfection in the

calibration.

In addition, we look at another class of biological eyes which have the same viewing directions. We study how the optic flow-fields can be used to facilitate the motion recovery in this eye arrangement, without resorting to stereopsis cues which will be too slow and heavy for animals with limited neural processing. Our algorithm proved to be on par with nonlinear BA at much less computational cost, a significant advantage in any visual system with a need for rapid visuomotor coordination.

List of Figures

List of Tables

Chapter 1

Introduction

Marr's computational vision paradigm has influenced deeply the development of computer vision. While Marr is correct in his observation that understanding the physical workings was not going to be enough and that we would also need to understand how the system was organized at a higher level (the computational question), it has led to a marginalization of the importance of the bodily aspect of the vision system. For although it is (probably) true to say that a computational understanding is in principle independent of the details of any specific implementation in hardware, the computational activities (especially for biological systems) are certainly heavily sculpted by the hardware implementation. With this in mind, we might ask ourselves if we have overlooked the wealth of organizational possibilities that are offered by the varieties of eye topography found in nature.

Contemporary research in robotics and AI viewed vision processing as the activity of an essentially situated agent: in particular, an agent that is at home in its proper bodily and environmental niche. It is likely to exploit just about any mixture of bodily and environmental resources along with significant interpenetration of perception, thought and

action. Yet the computer vision community has been primarily concerned with camera-type eyes that are frontally placed, forgetting that for many vertebrates, the eyes are laterally placed, not to mention the vast array of eye types that exist in the invertebrate world.

In this thesis, we look at the different topography of the eyes found in both vertebrates and invertebrates and see how the visual system can press maximal benefit from the opportunities afforded by a particular class of bodily realization that encompasses many animals. We are interested in leveraging the physical structure of this class of systems to solve Structure from Motion, a fundamental problem in computer vision.

1.1 Inspirations from Biological World

Theories about how insects exploit their compound eyes and the environment to carry out visuomotor tasks have been advanced indeed. For instance, in navigation, it was found that flying insects (Srinivasan and et al., 1991) are able to center their flight path in a corridor by balancing the image motion in their two eyes. In addition, honey bees have been shown to regulate flight speed by trying to keep the overall image motion as constant as possible (Srinivasan and et al., 1996). Such cooperation between bodily and environmental factors has also been implemented in various biomimetic approaches. The centering behavior of bees inspired the ”bee-bot” (Coombs and Roberts, 1993). By balancing the maximal flow on both sides, bee-bot centered its course between the nearest objects. While correcting its course, bee-bot’s camera actively counter-rotated to prevent the rotatory flow from contaminating the flow field. Other tasks investigated include altitude control landing (Srinivasan, Zhang, and Chahl, 2001) and view-based navigation

(Franz and et al, 1998a). However, no work exists on general ego-motion estimation that exploits the structure of the compound eye. The view seems to be that due to the limited neural resources of the insects, general ego-motion recovery is difficult; it is believed that only aspects of the ego-motion that are tailored to various visuomotor tasks are recovered. We show, however, that general ego-motion recovery can be achieved without resorting to complex and computationally expensive algorithms if we make use of the special arrangement of the compound eye.

Among vertebrates, the divorce of theoretical attention from the non-frontal eyes in the computer vision community is even more pronounced. Frontal eyes, where two eyes simultaneously gain very similar views of the same objects that lie in front of the head, as in humans, has received the most attention. However, in the great majority of vertebrates, each eye views a quite different part of the space that surrounds the head with various degrees of overlap of view between the two eyes. While there are computational works that look at a stereo pair or multiple cameras in simultaneous motion ((Baker, Ogale, and Fermüller, 2004; Tsao and et al., 1997; Zhang, 1995)), either with a fixed or varying epipolar geometry, their concerns are quite different. Stereo matching or feature tracking over multiple frames often plays an important role in such systems, and as such, corresponding points are needed. In our case of laterally placed eyes, correspondence of features between the two eyes is not even possible. Indeed, even in the insect eye system where the eyes are closely spaced, the visual field is so narrow that there might be little overlap between eyes. Thus the aforementioned multiple-cameras-in-motion models which require correspondences are not applicable. Are there, then, other affordances that exist in such an arrangement of multiple eyes, each covering different parts of the visual field, and each experiencing slightly different motions related to each other via

some rigid transformation? Do we just fuse the multiple inputs in a loosely coupled manner, that is, estimating the ego-motion of each camera independently and combining these ego-motions at the last stage? Or is there a tighter constraint at the lower level that allows a stable and preferably parsimonious solution for ego-motion recovery?

Biological systems in general seem to exploit the motion input from the different parts of the visual field. In vertebrates with laterally positioned eyes, such as rabbits and birds, as well as in arthropods equipped with panoramic vision, there are extensive spatial pooling of motion information and interactions between inputs from the opposite visual directions and they were shown to increase the sensitivity to particular types of optic flow field (e.g. rabbits: (Leonard, J.I.Simpson, and Graf, 1988); birds: (Wylie and Frost, 1999); bee: (Ibbotson, 1991); moth: (Kern, 1998), fly: (Haag and Borst, 2001)). Such exploitation of bodily factors results in parsimony of visual processing that is needed for integrated visuomotor coordination.

1.2 Motivation of the Thesis

Despite the biological evidence of such leveraging of motion information from opposite directions, it is certainly not the case that such a scheme is fully explored or understood computationally. The detailed mathematical aspect of the actual motion estimation is very much an open research question. As far as the authors are aware, save the works of (Lim and Barnes, 2007; Lim and Barnes, 2008; Thomas and Simoncelli, 1994) for the case of spherical cameras, there has been no work in the computational vision literature that looks at pooling optic flow inputs from opposite directions of the visual fields and investigate how such a pair of flows can be exploited in a tighter manner for ego-motion

estimation.

In this thesis, we pursue this strategy of pooling motion information from visual fields that are 180° opposite to each other for general ego-motion estimation. Such a strategy can be applied to a variety of vision systems, including the compound eye system and the system with laterally-positioned eyes. We show that such a pair of flows provide a measurement akin to parallax, and as in traditional parallax, it enhances information pickup for translation. This approach of pairing optic flows afforded by the physical arrangement of the cameras is in contrast to current works in computational vision, where the general motivation of having multiple cameras is to obtain feature matches across multiple viewpoints or to collect the optical flows from all the cameras, so as to resolve the inherent ambiguity in the ego-motion estimation problem. In that sense, the input from the multiple cameras is not fully exploited at the optic flow level in these conventional approaches; the redundancy enters the story only via the 3D relationships (i.e. the rigid transformations) that exist between the ego-motions experienced by each camera. The decoupling of the global translation and rotation is only initiated at the endpoint of a complex process which usually involves nonlinear estimation algorithm. Lastly, we also develop a mathematical understanding of our formulation in terms of its stability and robustness, and show how we have made the most of the information present in visual rays that are parallel but 180° opposite in direction.

The next section reviews the biological and the biomimetic robotics literature and discusses the wide variety of eye designs found in the vertebrates and the invertebrates.

1.3 Biological Vision Systems for Motion Estimation

1.3.1 Vertebrate Eyes: the case of birds

We review the visual systems of birds, the second largest group of vertebrates after fishes. Many bird species have laterally positioned eyes. Even in those cases where the eyes might appear to have binocular overlap of visual field (such as the raptors), the assumption that binocularity inevitably results in stereopsis has been questioned in birds ((McFadden, 1994), (Davies and Green, 1994)). It is pointed out in (Davies and Green, 1994) that stereopsis involves considerable neural processing and is too slow to control the estimation of distance and depth when a bird is landing upon a perch. The interactions between both eyes might indeed have to do with motion processing.

Interestingly, although many birds may appear to have significant overlap in the visual field, such direct estimation of the size of the frontal binocular field by causal examination of their appearance can be seriously erroneous. In particular, the width of frontal field binocularity is likely to be considerably overestimated. For instance, the Short-toed Snake-eagle eye optical systems produce a binocular field in the horizontal plane that appears 40° wide, but functionally the field is only half that width because there is no retina serving vision at the margin of the optical field. This is illustrated by Figure 1.1 from (Martin and Katzir, 1999).

This situation is not unique to this eagle and has been found in other species including Ostrich, herons and owls, and suggests that many bird species do not make full use of the potentially available binocular field. This finding casts doubts on the utility of stereopsis for general scene perception and locomotion, except for visuomotor tasks involving close objects such as the bills.

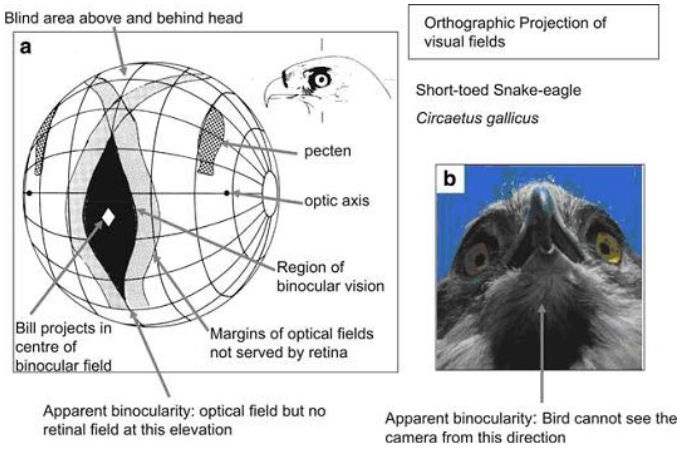


Figure 1.1: Depiction of visual fields for Short-toed Snake-eagle. **a)** A perspective view of an orthographic project of the visual field shows the binocular sector to the front of the head and the blind area above and the margins of the optical fields that are not served by retina. The projections of the optic axes, pectens and of the bill are also shown. It should be imagined that the bird is placed at the center of a transparent sphere that surrounds the head and the projections of the various features are drawn onto the surface, the orientation of the head is depicted in the inset drawing, but the median sagittal plane of the bird lies is in the same plane as the equator of the projection which is vertical and contains the projection of the bill. **b)** Photograph of a bird taken at a position in the sagittal plane below the bird that lies outside of the retinal visual field. The bird seems to have binocular vision at this point, as it is possible to see into the eye. However, there is no retina serving vision at this elevation and so the bird could not see the camera.

Interspecific comparison suggests that with respect to the features of the frontal, particularly the binocular field, three main types of visual field topography are found in birds. It is hypothesized that the characteristics of each particular type is determined primarily by feeding ecology, and secondarily by the requirements of provisioning young, rather than phylogeny or more general aspects of ecology and behavior.

• In species whose vision is used for the accurate control of bill position when pecking or lunging at prey, the maximum width of the binocular field lies between 20° and 30°. Examples of such birds are found in the eagles, herons, albatrosses, and hornbills.

• In species which do not use vision to guide bill or feet position in foraging, but rely upon tactile cues from the bill tip to locate items, or filter feed from surface waters, and do not provision their young, maximum binocular field width is only about 10°, and the binocular field may be only 5° wide at the horizontal.

• In the third type of visual field, the eyes are more frontally placed with wide binocular overlap (approximately 50°). However, it seems restricted to the owls and seems to be a unique case of eye topography in the avian kingdom.

The majority of these birds, despite their narrow binocular fields are capable of fast flight and maneuvering within both open and woodland habitats. This suggests that the control of flight in both open and complex woodland habitats does not require extensive frontal binocularity. Watching a sparrowhawk executing complex maneuvers to pursue agile preys through dense foliage tends to underscore its amazing visuomotor coordination, possibly achieved with only motion cues.

1.3.2 Invertebrate Eyes

Invertebrates have the greatest variety of eye types, with both camera eyes (e.g. *Cephalopods*) and compound eyes. We focus in this section on the compound eyes, the most abundant eye design in the animal kingdom, and probably its most adaptable. The first remarkable fact about its adaptability is that the layout of ommatidia is often matched to the spatial layout of the habitat, with higher concentration of ommatidia in some region termed as the acute zone. The presence of such acute zone implies that one region of the visual world is more important to the animals than the others. Such matching is clearly seen in the apposition eyes of animals adapted for life in a flat habitat such as a water strider that skates on a pond. All experience of a bright world is dominated by the horizon, and objects of interest occur at or very near the horizon. Thus they all possess eyes having an elongated horizontal region of enhanced resolution known as a visual streak. On the other hand, blowflies and hoverflies have a frontal-dorsal acute zone known as love spots used by the males to keep sight of females during high-speed pursuits. This

is a remarkable demonstration of the kind of close interaction between the vision system and the environment. Even the morphology of the compound eye itself adapts to the environmental surroundings and actions! There are many other aspects of the eye optics that are adapted to the environment (e.g. (Land and Nilsson, 2006), (Warrant, 2004)). Clearly these different designs confer special properties to the visual systems.

As far as motion processing is concerned, besides the kind of processing dedicated to specific reflexive responses such as collision avoidance and landing, general ego-motion recovery is understood to benefit from the spherical field of view of the compound eyes. Optic flow at positions that are 180° apart on a connecting meridian allows disambiguation of translation and rotation. During forward translation the optic flow across both eyes is directed backward. In contrast, during a pure rotation about the animal's vertical axis, optic flow is directed backward across one eye, but forward across the other eye. Such a strategy of pooling optic flow from both eyes appears to be adopted generally by arthropods. For instance, crabs use interactions between movement detectors that "look" in opposite directions of the visual field ((Blanke, Nalbach, and Varju, 1997)). Such interactions between two eyes are also found in many insects (e.g, bee: (Ibbotson, 1991); moth: (Kern, 1998), fly: (Haag and Borst, 2001)). For instance, the well-studied 'HS' cells of the fly estimate rotation from optic flow and is used to provide information for optomotor control and trajectory stabilization.

The thesis proposes a mathematical modeling of such motion processing and show that such a pair of flows, resulting from either lateral placement of vertebrate eyes or from compound eyes, provide a measurement akin to parallax, and as in traditional parallax, it enhances information pickup for translation. This is in contrast to current works in computational vision, where the general motivations of having multiple cameras have

been focused upon how a wide field of view can disambiguate the inherent ambiguity in the ego-motion estimation problem. The input from the multiple cameras is not exploited at the optic flow level; the redundancy enters the story only via the 3D relationships (i.e. the rigid transformations) that exist between the ego-motions experienced by each camera. The decoupling of translation and rotation is only initiated at the endpoint of a complex process which involves nonlinear estimation algorithm.

1.4 Thesis Organization

The organization of this thesis is as follows. Chapter 2 starts with the definition of Structure from Motion (S*f*M) and difficulties S*f*M faces, then it seeks to situate our proposed method in the vast S*f*M literature, relating our work to other research efforts and paradigms. Chapter 3 first briefly discusses the contribution of our linear quasi-parallax approach and lists those systems to which our approach applies. Then it introduces our quasi-parallax formulation in details, using first a basic setup with a single pair of cameras with opposing visual fields, and then an extended version that handles multiple pairs. To improve the robustness and accuracy of our linear algorithms, a Total Least Squares approach with an appropriate normalization scheme is presented. Then we go on to study the inherent ambiguity in our formulation using Fisher Information matrix and compares it against the "gold standard" solution obtained by the Bundle Adjustment method. Chapter 4 proposes two variant formulations of the linear quasi-parallax approach and the corresponding applicable biological eye combinations. Chapter 5 reports a set of experiments on motion recovery that were conducted using both "realistic" scenes adapted from range image input, and visual images of real scenes. Our method

was fully evaluated under different scenarios, and its performance was compared against the Bundle Adjustment algorithm. Chapter 6 carries out another set of experiments to study the performance of depth reconstruction in our method and Bundle adjustment. Finally this paper ends with discussion and conclusion in Chapter 7.

Chapter 2

Review of S*f*M Computational Literature

2.1 The Problem of Structure from Motion

2.1.1 Definition

The problem of structure from motion is to recover the 3D ego-motion of the camera and subsequently the 3D world structure from image sequences. Existing methods typically fall into two categories. One is the discrete case where we have a set of widely separated views obtained from multiple camera positions. It includes the classical eight-point algorithm (Hartley, 1997) and multiple view approaches (multiple-view tensor (Hartley and Zisserman, 2000). Discrete methods are based on image correspondences, which are two points in two the images projected from the same 3D world point. Given a set of image correspondences, discrete methods will estimate an essential matrix, which can later be decomposed into 3D camera translation and rotation.

The other category is known as the differential case. We have a monocular camera in motion and a sequence of densely sampled image velocities (optical flows) acquired under a relative instantaneous motion between the camera and the scene. This thesis adopts the differential approach.

We adopt a pinhole camera model with perspective projection. A point $\boldsymbol{P}$ in the world projects onto a point $\boldsymbol{p}$ in the image plane which is f pixels away from the optical center. We have $\boldsymbol{p} = f\frac{\boldsymbol{P}}{Z}$. Here we assume the focal length f is known in a calibrated camera.

Denote the camera is undergoing a global translation $(U, V, W)^T$ and a global rotation $(\alpha, \beta, \gamma)^T$. Assuming (u, v) is the optical flow at image point (x, y) arising from a scene point with depth Z, we have (Longuet-Higgins and Prazdny, 1980):

$$\begin{aligned} u &= \frac{u^{tr}}{Z} + u^{rot} = \frac{Wx - fU}{Z} + \frac{\alpha xy}{f} - \beta(\frac{x^2}{f} + f) + \gamma y \\ v &= \frac{v^{tr}}{Z} + v^{rot} = \frac{Wy - fV}{Z} + \alpha(\frac{y^2}{f} + f) - \beta\frac{xy}{f} - \gamma x \end{aligned} \tag{2.1}$$

where $\frac{1}{Z}(u^{tr}, v^{tr})$ and (u^{rot}, v^{rot}) are the components of the flow due to the translation and the rotation respectively. $u^{tr} = (Wx - fU), v^{tr} = (Wy - fV)$ and $u^{rot} = \frac{\alpha xy}{f} - \beta(\frac{x^2}{f} + f) + \gamma y, v^{rot} = \frac{Wy - fV}{Z} + \alpha(\frac{y^2}{f} + f) - \beta\frac{xy}{f} - \gamma x$. Canceling the depth Z from the above two equations gives us the differential epipolar constraint:

$$uv^{tr} - vu^{tr} = u^{rot}v^{tr} - v^{rot}u^{tr} \tag{2.2}$$

Fully expanding Equation (2.2) yields many nonlinear terms on the right-hand side, most of which are the coupling terms between translation and rotation generated by the products $u^{rot}v^{tr}$ and $v^{rot}u^{tr}$. This coupling contributes to the formation of the bas-relief valley, which will be discussed in the next section.

2.2 Review of Computational Literature

2.2.1 Two-view Motion Estimation Methods

Ego-motion estimation methods typically are divided into two categories. The first is the discrete case where the camera displacement and the resulting 3D world structure can be recovered from two or multiple widely separated views of the same scene, as shown in Fig 2.1.

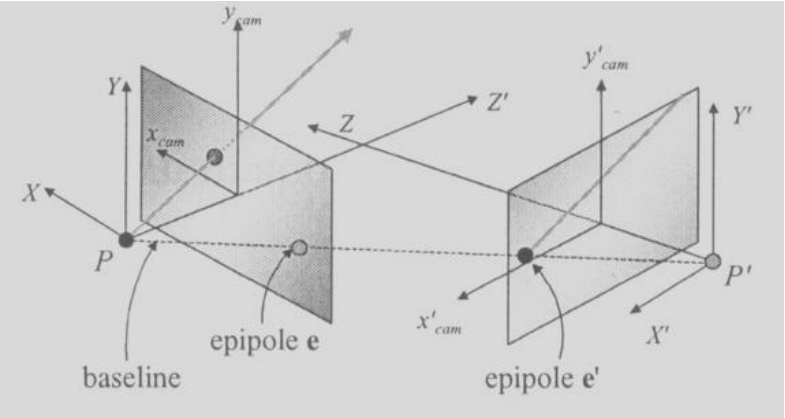


Figure 2.1: General stereo imaging configuration

These views may be acquired simultaneously as in a stereo rig, or acquired sequentially, such as by a camera moving relative to the scene. One of the prominent approaches is the essential matrix approach (Toscani and Faugeras, 1986). Given image point correspondences, we enforce the intrinsic geometric constraint (termed as the epipolar constraint) from which a so-called essential matrix is recovered. For the two-view discrete case, some classical linear methods include the eight-point algorithm which is based on SVD-decomposition to decompose the essential matrix into 3D camera translation and rotation (Hartley, 1997). A general principle for such discrete, epipolar-constraint based algorithms to work well is that the displacement (especially the translation, also known as baseline) between the two images needs to be relatively large. Otherwise, due to the

small translation this algorithm becomes singular and the depth reconstruction becomes less reliable too. Unfortunately, in the case of a large baseline, the search distance for the corresponding point in the other view generally is also large. Hence feature correspondence is difficult to establish.

The other class of motion estimation algorithms is the differential case: the camera motion and the 3D world are recovered from a sequence of densely sampled image velocities (also known as optical flow) acquired under a relative motion between the monocular camera and the scene as in Fig 2.2.

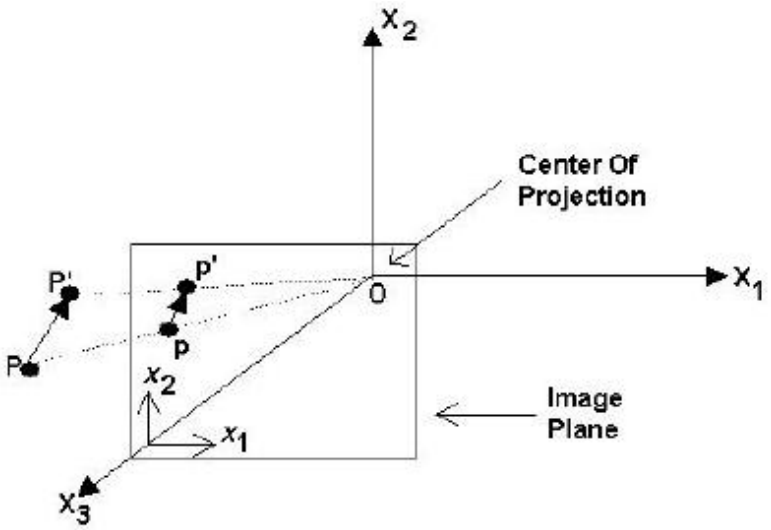


Figure 2.2: 3D motion field projected onto the image plane: 2D motion field

Most algorithms in this category are based on the differential epipolar constraint which relates optical flow to the camera translation and rotation. The camera motion can be solved either through numerical optimization or by linear methods after a cost function is obtained from the epipolar constraint. Linear subspace method proposed in (Heeger and Jepson, 1992) is one such example. One big advantage of the differential approach is that there is no correspondence problem. But the computation of optical flow involves assumptions that might be violated, for example in the aperture problem

(Trucco and Verri, 1998) illustrated in Fig 2.3. In the image the solid line moves to the dash line. If we view the image through a small aperture, we cannot tell where this point (x_i, y_i) moves to due to the lack of differently oriented edges and distinctive points. Thus we do not know the image velocity $\vec{u}$ and only the component perpendicular to the edge $\vec{u_n}$ can be computed. This is known as the aperture problem.

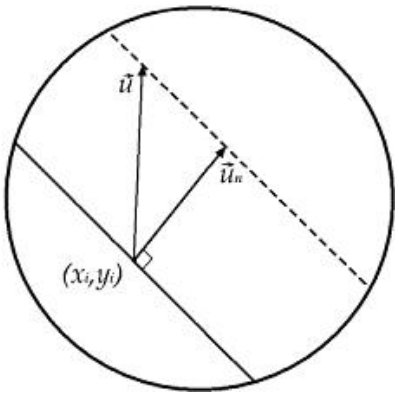


Figure 2.3: Illustration of the aperture problem

It is known that linear two-view methods yield biased motion estimates and many nonlinear methods have been proposed to overcome this problem. Among various methods, the bundle adjustment (Triggs and et al., 2000) (referred hereafter as BA) stands out with its optimal performance and is regarded as the "Gold Standard". By minimizing the reprojection error between the measured flows and the estimated flows, BA yields maximum likelihood estimates given a Gaussian image noise.

One thing to note is that in both discrete and differential classes, the two-view methods suffer from the bas-relief ambiguity arising from the coupling between the translation and rotation. This bas-relief ambiguity is caused by the geometry of the problem itself and thus cannot be removed by any statistical schemes. We will discuss this in details in the next section.

2.2.2 Bas-relief Ambiguity

According to a number of geometrical investigations of the S*f*M behavior (Fermüller and Aloimonos, 2000; Xiang and Cheong, 2003), S*f*M problem suffers from inherent ambiguity even in the noiseless case. And some of the well-known visual confusions can be attributed to this ambiguity. For example, in a small field of view, the optical flow field of a rotation around the vertical y-axis is very similar to that of a translation along the horizontal x-axis. According to (Xiang and Cheong, 2003) paper, when the FOV=28^o, the ambiguity is already very prominent.

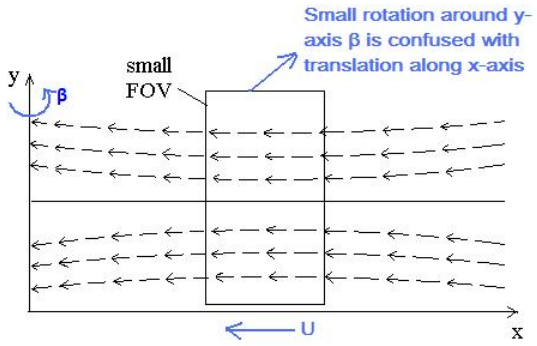


Figure 2.4: Confusion between translation in the x-axis and rotation around the y-axis.

The analysis of the error surface topology shows that the coupling of the translation with rotation generates a so-called bas-relief valley in many motion-scene configurations. Figure 2.5 from (Fermüller and Aloimonos, 2000) shows that the valley is quite flat and consequently many motion candidates could serve as the minimum. Thus when the algorithm is minimizing the error function, it is very likely that the minimum returned is not the true solution, thus introducing errors into the estimates.

As we mentioned above, due to the coupling between the translation and rotation, the error residue caused by an error in the translational estimate can be compensated by suitable choice of error in the rotational estimate. Since this ambiguity is caused by the

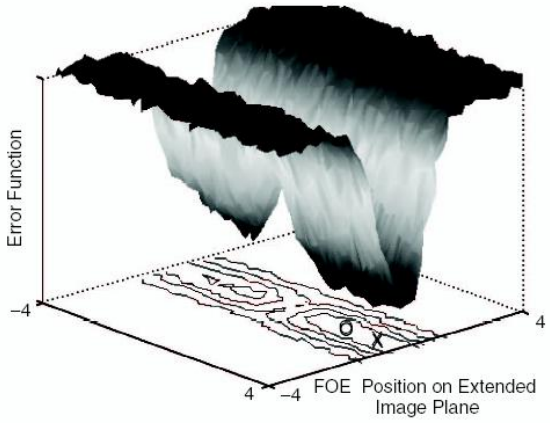


Figure 2.5: Topology of the cost function surface in the space of the direction of translation where 'O' is the true FOE position and 'X' is the solution of the optimization algorithm.

geometry of the problem itself, it cannot be removed by any statistical schemes.

Taking a closer look at the error profile along the bas-relief valley, we see that the location of the true FOE, i.e. the direction of the true translation, has a critical influence on the shape of the bas-relief valley. In general, largely lateral translations present a more difficult scenario for SfM recovery as the location of the so-called opposite minimum will approach infinity with its residual value approaching zero (Xiang and Cheong, 2003). Hence in Fig 2.6, a large part of the bas-relief valley becomes very flat, thus resulting in a highly ambiguous situation. In contrast, in the pure forward motion case, the opposite minimum merges with the true solution at the origin; thus the local minimum disappears and the true solution clearly stands out.

Correspondingly, according to the error analysis of SfM in the discrete case, there is also an intrinsic bas-relief ambiguity independent of the choice of objective cost functions (Ma, Kosecka, and Sastry, 2001; Adiv, 1989; Weng, Ahuja, and Huang, 1993). It is shown in the small field of view, for each lateral translation (parallel to the image plane), there exists a corresponding type of rotation such that the displacement field of a translation can be interpreted as a rotation. Even small pixel-level errors will cause large

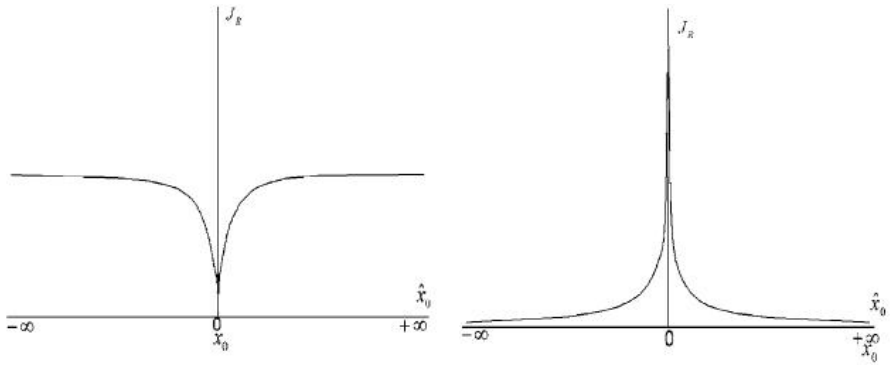


Figure 2.6: Error profiles of the bas-relief valley in the limiting cases. J_R is the cost function. x_0 is the x-coordinate of the true FOE while $\hat{x_0}$ is that of the estimated FOE. (a) pure forward motion where the true FOE is at the origin and (b) pure lateral motion where the true FOE lies at infinity.

errors in the estimated motion.

2.2.3 Multiple-view Motion Estimation Method

Multiple-frame algorithms such as multiple view tensor (Hartley and Zisserman, 2000)) incorporate redundancy from multiple frames to overcome the inherent bas-relief ambiguity faced by two-view methods. The extension in space results in a geometry described by the so-called multiple-view tensor, which has analogous properties to the fundamental matrix of two-view geometry, but enjoys additional benefits over two-view geometry. Take the three-view case as an example in Fig 2.7. Given the cameras, in the two-view case each point correspondence provided four equations on the three degrees of freedom (the position) of the 3D point. In three views there are six equations on the three degrees of freedom. So with more equations, the measurement errors in the input can be minimized and the solution becomes more stable.

However, this approach has its limitation. As we have to establish feature correspondence across all the frames, the difficulty in correspondence compounded. Moreover, we still face this dilemma: use large baselines to gain robustness or use small baselines to

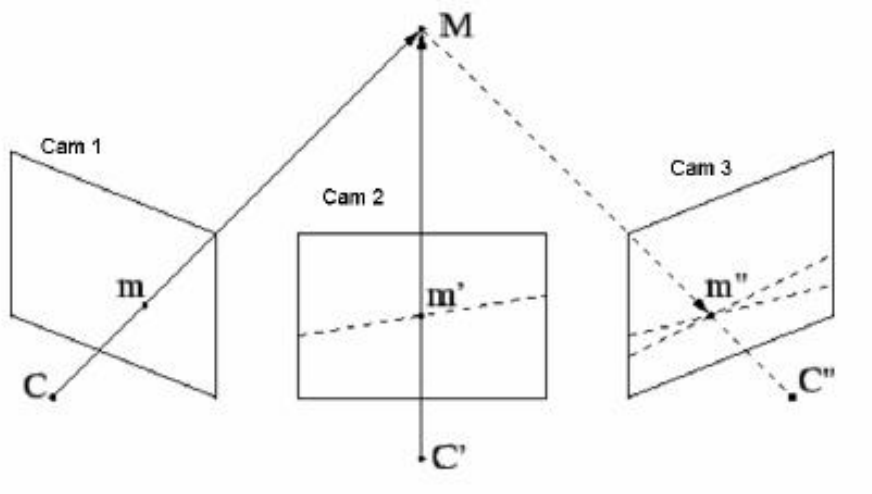


Figure 2.7: Three-view geometry.

ease correspondence problem.

Analogously, there is also an extension in time of the two-frame optical flow method, one example of which is the factorization method (Tomasi and Kanade, 1992). Based on the assumption that the camera model is orthographic (objects are distant with respect to their size), it tracks P scene points, not all coplanar, across F image frames with F $\geq$ 3. Constructing a 2F $\times$ P measurement matrix, the authors then use SVD technique to factor the matrix into two matrices which describe object shape (3D scene points' co-ordinates) and camera motion respectively. This method works fast and yields accurate results. However, because of the orthographic assumption, the component of the translation along the optical axis cannot be determined. Moreover, this approach no longer works in domains where perspective camera effect cannot be ignored.

Another class of algorithms combine stereo and motion analysis to solve SFM so that the two visual cues complement each other (Okutomi and Kanade, 1993; Baker, Ogale, and Fermüller, 2004). These works use two cameras to take a stereo pair of image sequences to facilitate the establishment of stereo correspondences. Afterwards, motion cue is ignored and only stereo cue is used to solve for motion parameters and later for

3D reconstruction. Thus the fusions in these works are characterized by loose coupling. The focus of our work, however, is on tightly integrating the two cues so that both ease of correspondence and accurate recovery of motion parameters can be achieved. The joint information from both cues is utilized throughout the whole recovery process.

2.2.4 Motion Parallax

Traditional parallax method (Longuet-Higgins and Prazdny, 1980) is a different approach to the SfM problem. It separates the rotational flows from the translational flows by observing that for near-coincident image points across a depth discontinuity, the difference in their flows cancels the rotational flow. Provided with sufficient number of such points, this method is able to solve for the translation, after which the rotation can be readily obtained. Unfortunately it is very difficult to find enough nearby points with sufficiently large depth differences, and thus its usefulness is limited. Alternative approaches called "plane+parallax" have been proposed (Anandan and Irani, 2002). They assume a dominant plane in the scene or a piecewise planar world model. The alignment with respect to this plane removes the rotation and leads to an epipolar motion field (or a parallax field), from which the ego-motion can be computed. However for a general scene with arbitrary structure, the planar assumption might be violated.

2.2.5 Non-conventional Cameras

In recent years, camera cluster in various configurations has seen increasing popularity thanks to falling costs and miniaturization. In tandem with this development, there have been several theoretical analysis analyzing the properties of such camera clusters. Pless

(Pless, 2004) obtained a generalized epipolar constraint for multi-camera setup and compared the fitness of several designs in resolving the rotation-translation ambiguity via the Fisher information matrix. Sturm (Sturm, 2005) considered a general imaging model that incorporates multiple-camera views and analyzed the geometry of the problem. The minimal information needed to solve the motion recovery problem for a multiple-camera platform was discussed in (Stewenius and Astrom, 2004). Both (Stewenius and Astrom, 2004) and (Sturm, 2005) assumed that correspondences were available, but as discussed in the preceding sections, correspondences are difficult and computationally expensive to obtain. The configuration of these multiple cameras is in general arbitrary, and thus the algorithms proposed are also general, independent of any specific camera arrangement.

One of the works exploiting such multiple-camera system in simultaneous motion is the Argus eye (Baker, Ogale, and Fermüller, 2004) which consists of nine outward-pointing cameras. For each camera, a set of camera motion candidates with the smallest residual errors were found. The intersection of these candidates when expressed in the global coordinates was taken as the global motion. The fusion of information from the various cameras in the Argus Eye can be regarded as a loose form of couplings (in the sense of (Clark and Yuille, 1994)), since the individual motion estimates are computed independently and the fusion takes place only at the last stage.

In the work of (Tsao and et al., 1997), the epipolar constraints of each camera are collected together, and a nonlinear residual function is obtained when the individual camera motions are expressed in terms of the global motion. This nonlinear function was then minimized to obtain the global motion. Similar to works reviewed in the preceding paragraphs, the multiple camera model analyzed is a general one, without any attempt to exploit the constraints afforded by a particular configuration of the multiple-camera

setup.

Neumann et al (Neumann and et al., 2004) proposed a linear plenoptic approach for motion estimation. Our work is similar to this work in the sense that both methods are linearly formulated and do not need correspondence. Like (Neumann and et al., 2004), our work can be applied to a compound eye system, but our formulation can also be applied to other eye topology. There are, however, several major differences. First, our linear method stems from considering projection rays that are parallel but opposite in direction. In contrast, the linearity of the plenoptic method comes from a five-dimensional plenoptic function, which is unmeasurable in a conventional camera. Due to this difference, our system can be built from conventional pinhole cameras, whereas the plenoptic method requires a specially designed sensor whose physical structure is currently unrealizable. Secondly, in our method, the translation is recovered from a quasi-parallax term where information pick-up for translation is enhanced; whereas the rotation is estimated in a separate post-translation step. In this way, the coupling between the translation and the rotation is minimized, since only those measurements best suited for translation recovery (and rotation recovery respectively) are utilized. In contrast, the plenoptic approach solves for both the translation and the rotation simultaneously. Though the motion recovery enjoys the benefit of a spherical FOV, it is not clear if the plenoptic approach is the best way of removing the coupling between the translation and the rotation.

Other pertinent works are that of (Lim and Barnes, 2007; Lim and Barnes, 2008; Thomas and Simoncelli, 1994) which estimate the epipole in a spherical eye. These works are related to ours in so far that they exploit the information present in an opposite pair of visual rays, which are termed as an antipodal pair. However, such a camera

system with a single viewpoint is a serious qualification as far as modeling biological visual systems is concerned, as most natural visual systems such as the lateral eyes of vertebrates and the insect compound eyes are not of a single viewpoint. For instance, in a compound eye system, each ommatidium has its own optical center, situated on different parts of the substrate surface or the head. Our paper examines fully the computational implications of such eye arrangements found prevalently in the animal kingdom. Furthermore, we use Fisher information matrix to explicitly characterize the inherent ambiguity in such eye arrangements, where each eye might only have a small field of view. In comparison, the methods of (Lim and Barnes, 2007; Lim and Barnes, 2008; Thomas and Simoncelli, 1994) have little to say about these numerical aspects, since they start with a spherical field of view, which hardly suffers from the bas-relief ambiguity. Thus, it is unclear whether their good performances benefit from the formulation, or simply result from the spherical field of view. For instance, in (Lim and Barnes, 2007), if all the antipodal pairs are closely clustered, the great circles in their formulation would span a small angle in their orientations; thus the intersections of these great circles might not be well-localized enough for accurate epipole estimation. Where applicable, further differences between (Lim and Barnes, 2007; Lim and Barnes, 2008; Thomas and Simoncelli, 1994) and our work will be highlighted in later sections.

Chapter 3

Linear S*f*M based on Quasi-Parallax

This thesis proposes a method for solving ego-motion and applies to systems where there exists visual rays that are parallel to each other but facing the opposite direction, and the entire system moves in tandem. These visual systems can be realized in a variety of ways such as via the compound eye or via the laterally placed pinhole cameras (for details, see Section 3.2). We show that our formulation presses maximum advantage from the physical structure of these visual systems, while conventional methods based on some nonlinear constraints do not have the optimal way of combining information and thus still suffer from the bas-relief ambiguity to some extent, even for the gold-standard BA technique.

3.1 Our Contribution

The main contribution of our work lies in proposing a method that utilizes different and appropriate ensembles of visual rays for estimating translation and rotation respectively. For translation recovery, we consider visual rays that are parallel but facing the opposite

direction. We call such pairs of visual rays the matching pairs. In a geometrical sense, this grouping operation that we perform is akin to that in obtaining the parallax, as the matching pair have identical rotational flows. However, we term the resulting difference in motion flows as quasi-parallax, because it still contains weak induced translational terms caused by the global rotation, and thus carries terms determined by the rotational parameters. Despite this, the effect of the induced translational terms has been largely reduced and thus the global translation can be accurately recovered. For the same reason, such quasi-parallax term is clearly not suitable for rotation recovery. Instead, the rotation is computed in a post-translation step, by looking at the individual epipolar constraint of each camera. In this way, we are able to press maximum advantage from the diametrically opposite visual field available in such visual system.

In the terminology of the plenoptic function, we are using parallel rays for translation recovery and a pencil of rays for rotation recovery. Thanks to this two-staged recovery process, the translation can be recovered well even with a dominant rotational flow; a good translation estimate in turn benefits the rotation recovery. This is in contrast to those nonlinear computationally expensive methods, which use the same input to estimate the translation and the rotation together, and typically involve heavy optimization over six motion parameters simultaneously (Pless, 2004; Triggs and et al., 2000). In consequence, the estimate error of the translation and that of the rotation are intertwined and affect each other. This undesirable coupling effect is especially obvious when either the translational flow or the rotational flow is dominant.

To quantify the relative merits of our linear method and the nonlinear BA, we use Fisher information matrix as a tool to analyze the inherent ambiguity in the two formulations. We show that our formulation has a more optimal integration of the information

from multiple cameras, in the sense that it resolves the bas-relief ambiguity much better, especially under small field of view.

With regards to numerical implementation, we adopt data normalization and the Total Least Squares approach (TLS) so that the system is well-conditioned and robust to noise perturbation. Compared to Linear Least Squares, TLS is more robust for our *errors-in-variables* system. Our quasi-parallax method also improves the feasibility of the traditional parallax idea: we only require a matching pair of points in two cameras to have some depth difference, instead of requiring two coincident points in the same camera to have depth difference. Clearly, this assumption is much more readily satisfiable, considering that each camera faces opposite directions and is likely to view scenes with different depths.

There are other advantages to our formulation. Like the traditional parallax method, our method is based on optical flows; thus no stereo correspondence is needed. Another important advantage of our algorithm is that the length of the baseline (the distance between the camera centers) is not necessary for the motion estimation, as we show in the next section. This advantage renders our method suitable for cases where the baseline might be difficult to measure.

In sum, the contribution of our work lies in the following. First, we analyze the aptitude of a class of eyes (possessing the common characteristics of having matching pairs of visual rays) for SfM. Then a linear algorithm is put forth to address both the geometrical and numerical difficulties associated with the motion estimation problem. Geometrically, we are able to make full use of the inherent robustness of such wide-FOV camera clusters by tightly coupling the information from each individual camera at the flow level. We make use of the quasi-parallax to accurately recover the translation

by weakening the rotation. The robustness of the algorithm is enhanced as we eschew correspondences and do not require strict parallax. Numerically, with appropriate data normalization and the TLS approach, the linearization proves to behave well, with both the translation and rotation being recovered well with very little biases. Compared to nonlinear methods, our method runs at a much less computational cost and is much faster.

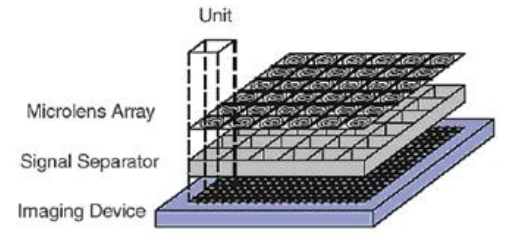
3.2 Systems to which our method can be applied

During the last 20 years, high-speed micro-circuitry, sensor arrays, arrays of micro lenses, gradient index materials, special glasses, and plastic optics manufacturing techniques are starting to give us more suitable building blocks for artificial eyes arranged in different configurations.

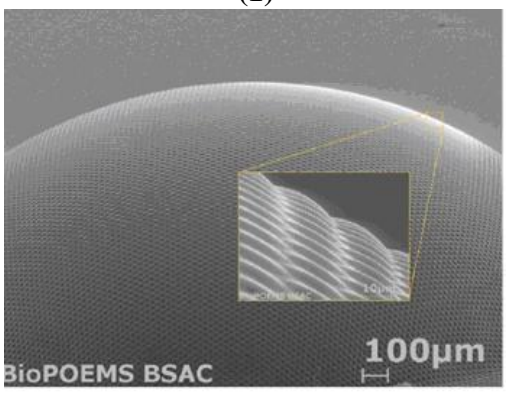
3.2.1 Artificial Compound Eye

Biologically inspired optical science is a relatively new and expanding field. The recent development of reconfigurable soft lithography using polydimethylsiloxane (PDMS) allows the creation of unconventional three-dimensional (3D) polymeric optical systems similar to biological ones, which are themselves constructed from biological polymers.

New micro-machining technologies have miniaturized these devices to a scale never before possible. Ogata et al. (Ogata, Ishida, and Sasano, 1994) fabricated an artificial compound eye and integrated retina in the 1990s using planar arrays of gradient refractive index (GRIN) rods to focus light through pinholes onto a photodetector array. Their resolution was only 16 by 16 but has led to modern versions that have the ability to



(1)



(2)

Figure 3.1: **(1)** Microlens array from Ogata et al **(2)** An artificial compound eye fabricated by the biologically inspired 3D optical synthesis method from Jeong et al

capture full color images (Duparre and et al.,). Even newer examples of apposition eyes have ommatidia arranged normal to a sphere, more faithful to their natural counterparts. Hornsey et al. (Hornsey and et al., 2004) constructed an optic dome covered with eyelet lenses made of glass fiber bundles. The first artificial ommatidia by self-aligned microlenses and waveguides were created by (Kim, Jeong, and Lee, 2005). This was followed by a 3D compound eye with self-aligned waveguides and individual microlens units on a spherical surface by (Jeong, Kim, and Lee, 2006). The ommatidia are arranged along a hemispherical polymer dome such that each points to a different direction, allowing for a wide field of view, similar to that of the natural eye. Neumann et al (Neumann and et al., 2004) also presented a design concept for a compound eye sensor and showed how it can be used to solve the ego-motion estimation problem.

3.2.2 Lateral Eyes

The visual system of lateral eyes is commonly found in vertebrates, especially birds. Located at opposite sides of the head, the two laterally positioned eyes have viewing directions that are parallel but 180° apart. Each eye works independently and views different parts of the world with non-overlapping visual fields. A special case arises when the two optical centers coincide, reducing the lateral eye system to a panoramic eye configuration. Any visual system configured in this fashion can obtain matching rays in the entire or part of the visual field of each camera, and it has the virtue of simplicity in construction.

3.2.3 Conical Mirror

The following conical mirror camera from (Chahl and Srinivasan, 1997) has become quite popular in biomimetic vision systems since it provides a two-dimensional model of the almost omnidirectional insect eye, and it is relatively easy to construct (see also (Franz and et al, 1998b),(Huber and Bülthoff, 1998)). Referring to Figure 3.2 if the upper limiting ray l_u is above the horizon (this is the case if $\alpha > 90^\circ$ and $R > -h\cos\alpha\tan 0.5\alpha$), then matching pairs of visual rays can be found. In particular, if we let $\theta(l)$ denote the angle made by the visual ray l with the horizon, then those visual rays l_i with $|\theta(l_i)| < \min(|\theta(l_1)|, |\theta(l_u)|)$ and which are mapped inside the circular disk with radius $\rho_{\max}$ in the image plane would have matching pairs.

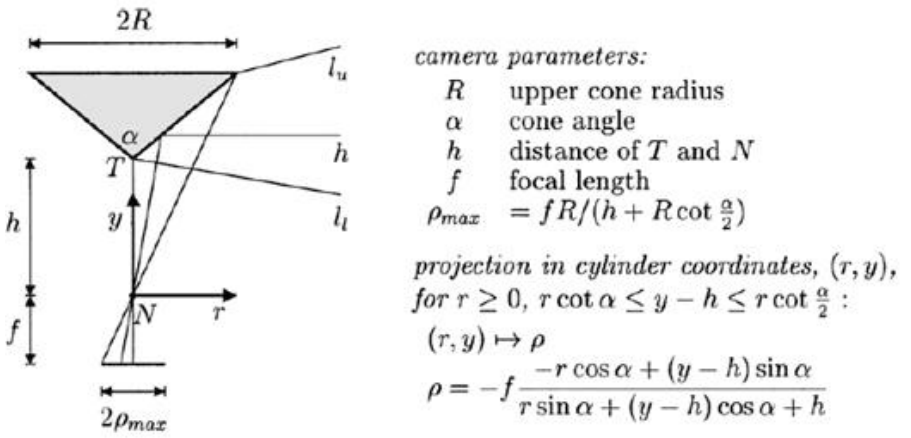


Figure 3.2: Imaging geometry for a conical mirror camera. N, camera nodal point; T, tip of conical mirror; l_u, upper limiting ray; h, horizontal ray; l_1, lower limiting ray. The conical mirror camera allows for capturing omnidirectional images without rotating a camera. A ring-shaped visual field between l_1 and l_u is mapped to a circular disk with radius ρ_{max} in the image plane. The visual field contains the horizon if $\alpha > 90°$ and $R > -h\cos\alpha\tan 0.5\alpha$

3.3 Technical Details of Our Approach

3.3.1 Prerequisites

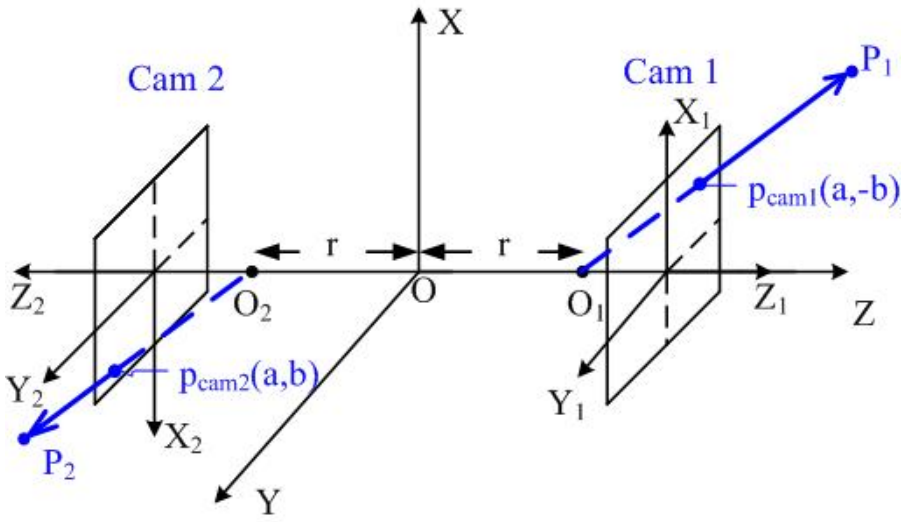


Figure 3.3: Configuration of the laterally-placed pair of cameras. p_{cam1} and p_{cam2} form a matching pair of points.

Figure 3.3 shows the basic set-up of our system, configured as a laterally placed pair of cameras. An extended set-up involving more than two cameras is studied in Section 3.3.3.

Here, the two cameras face opposite directions and their optical centers are of equal distance (termed as radius r) from the global origin O. The two cameras are attached rigidly to the body and move together according to a global motion $(\mathbf{v}, \omega)$ expressed in the world coordinate system $\{$O-XYZ$\}$. The c^{th} camera's translation $\mathbf{v}_c$ and rotation ω_c are related to the global motion by the following respectively:

$$\mathbf{v}_c = \mathbf{R}_c^T \cdot (\omega \times \mathbf{T}_c + \mathbf{v}), \quad \omega_c = \mathbf{R}_c^T \cdot \omega \tag{3.1}$$

where $\mathbf{R}_c$ and $\mathbf{T}_c$ denote the orientation and translation of the c^{th} camera relative to the world reference frame respectively:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{T}_1 = \begin{bmatrix} 0 \\ 0 \\ r \end{bmatrix}, \mathbf{R}_2 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{T}_2 = \begin{bmatrix} 0 \\ 0 \\ -r \end{bmatrix}$$

Assuming the global motion executed by the system is given by translation $\mathbf{v} = (U, V, W)^T$ and rotation ω=$(\alpha, \beta, \gamma)^T$, the individual 3D motions of cameras are:

$$\begin{aligned} \mathbf{v}_1 &= (U + \beta r, V - \alpha r, W)^T, \quad \omega_1 = (\alpha, \beta, \gamma)^T \\ \mathbf{v}_2 &= (-U + \beta r, V + \alpha r, -W)^T, \omega_2 = (-\alpha, \beta, -\gamma)^T \end{aligned} \tag{3.2}$$

This paper adopts a perspective pinhole camera model with known focal length f. Assuming (u, v) is the optical flow at image point (x, y) arising from a scene point with depth Z, we have:

$$\begin{aligned} u &= \frac{u^{tr}}{Z} + u^{rot} = \frac{Wx - fU}{Z} + \frac{\alpha xy}{f} - \beta(\frac{x^2}{f} + f) + \gamma y \\ v &= \frac{v^{tr}}{Z} + v^{rot} = \frac{Wy - fV}{Z} + \alpha(\frac{y^2}{f} + f) - \beta\frac{xy}{f} - \gamma x \end{aligned} \tag{3.3}$$

where $\frac{1}{Z}(u^{tr}, v^{tr})$ and (u^{rot}, v^{rot}) are the components of the flow due to the translation and the rotation respectively. Canceling the depth Z from the above two equations gives

us the differential epipolar constraint:

$$uv^{tr} - vu^{tr} = u^{rot}v^{tr} - v^{rot}u^{tr} \tag{3.4}$$

Fully expanding Equation (3.4) yields many nonlinear terms on the right-hand side, most of which are the coupling terms between translation and rotation generated by the products $u^{rot}v^{tr}$ and $v^{rot}u^{tr}$. This coupling contributes to the formation of the bas-relief valley under small FOV: the residue caused by error in the translational estimate can be compensated by suitable choice of error in the rotational estimate.

3.3.2 The Basic Two-stage Recovery Algorithm

Our estimation algorithm consists of two stages where the translation and the rotation are recovered separately. In the first stage, the translation is estimated from the quasi-parallax terms in which the translational flows are dominant. Given the translational estimate, the second stage constructs a linear system suitable for recovering rotation.

To obtain the quasi-parallax terms for translation recovery, we collect from the camera pair projection rays that are parallel but opposite in direction. Such pair of visual rays project onto the two image planes a pair of image points, which we term as the matching points. As in Figure 3.3, $\mathbf{p}_{cam1}$ and $\mathbf{p}_{cam2}$ are a pair of matching points. $\mathbf{p}_{cam1}$ is projected from visual ray $\mathrm{O_1P_1}$ and $\mathbf{p}_{cam2}$ is projected from $\mathrm{O_2P_2}$. If the image coordinate of $\mathbf{p}_{cam1}$ is $(a, -b)$, it is evident that $\mathbf{p}_{cam2}$ should lie at the position of (a, b). In general, we use $\mathbf{p}_{cam1} = (x^i, y^i)$ and $\mathbf{p}_{cam2} = (x^i, -y^i)$ to represent the i^{th} matching pair. In the $c^{th}(c = 1, 2)$ camera, denote (u_c^i, v_c^i) as its optical flow measured at the i^{th} matching pair. For notational convenience, we omit the index i where it is clear from the context, and thus $\frac{1}{Z}(u_c^{tr}, v_c^{tr})$ denotes the translational flow and (u_c^{rot}, v_c^{rot}) the rotational flow.

3.3.2.1 Stage 1: Recovering the global translation

At the matching points, we substitute the respective camera rotations in Equation (3.2) into Equation (3.3). It is clear that the rotational flows at the pair of matching points are identical in magnitude:

$$u_2^{rot} = u_1^{rot} \triangleq u^{rot},\ v_2^{rot} = -\,v_1^{rot} \triangleq -v^{rot} \tag{3.5}$$

With these two equalities in hand, we subtract the respective epipolar constraints (Equation (3.4)) of camera 1 and camera 2 and obtain:

$$u_1 v_1^{tr} - u_2 v_2^{tr} - v_1 u_1^{tr} + v_2 u_2^{tr} = u^{rot}(v_1^{tr} - v_2^{tr}) - v^{rot}(u_1^{tr} + u_2^{tr}) \tag{3.6}$$

Substituting the respective motions of camera 1 and camera 2 in Equation (3.2) into Equation (3.6), we obtain:

$$\begin{aligned}&(v_1 + v_2)U + (u_2 - u_1)V + \frac{(u_1 - u_2)y - (v_1 + v_2)x}{f}W \\ =&2r(u^{rot}\alpha + v^{rot}\beta) - (u_1 + u_2)\alpha r - (v_1 - v_2)\beta r\end{aligned} \tag{3.7}$$

This equation is linear in the translation unknowns (U, V, W), with all the coupling terms between $\mathbf{v}$ and ω eliminated. Note that if we set $r = 0$, the preceding would reduce to the following equation:

$$\frac{u_1 - u_2}{v_1 + v_2} = \frac{Wx - fU}{Wy - fV} \tag{3.8}$$

This is equivalent to the constraint obtained by summing the optical flows at antipodal points of a spherical camera in (Lim and Barnes, 2007) and then eliminating the depth-related factor K in their formulation. It is also related to the earlier work of (Thomas and Simoncelli, 1994) which takes the cross product of the optical flow with the position vector to obtain the angular flow, essentially a dual representation of the usual optical flow. In terms of the geometrical constraint utilized, there is no difference between the

formulations of (Thomas and Simoncelli, 1994) and (Lim and Barnes, 2007; Lim and Barnes, 2008) and the $r = 0$ special case in our formulation.[1]

Coming back to our system, normally it produces multiple pairs of matching points in the form of Equation (3.7). Collecting all the N equations from the entire set of matching points, we arrive at:

$$\mathbf{A}_{(N\times 3)}\mathbf{x}_1 = -r\mathbf{B}_{(N\times 6)}\mathbf{x}_2 \tag{3.9}$$

where $\mathbf{x}_1 = [U, V, W]^T$, $\mathbf{x}_2 = [\alpha, \beta, \alpha\beta, \beta\gamma, \alpha\gamma, \beta^2 - \alpha^2]^T$, and the corresponding i^{th} row of $\mathbf{A}$ and $\mathbf{B}$ are as follows:

$$\begin{aligned}
\mathbf{a}_i &= [v_1^i + v_2^i,\ u_2^i - u_1^i,\ \frac{(u_1^i - u_2^i)y^i}{f} - \frac{(v_1^i + v_2^i)x^i}{f}] \\
\mathbf{b}_i &= [u_2^i + u_1^i,\ v_1^i - v_2^i,\ 2\frac{x^{i2} - y^{i2}}{f},\ 2x^i,\ -2y^i,\ 2\frac{x^i y^i}{f}]
\end{aligned} \tag{3.10}$$

The right-hand-side (RHS) of Equation (4.12) or 3.9 still contain terms in α, β, γ and r; they enter the equation via the induced terms $\omega \times \mathbf{T}_c$ in the individual camera translations $\mathbf{v}_c$. That is, they arise from the translation induced by the global rotation ω. In this sense, the term $r\|\mathbf{B}\mathbf{x}_2\|$ on the RHS of Equation (3.9) can be viewed as a residue, resulting from imperfect parallax arising from the induced translation $\omega \times \mathbf{T}_c$. Clearly, if the radius r is zero, no induced translation exists, and we will obtain perfect parallax.

In most scenarios, this residue caused by the induced translation $\omega \times \mathbf{T}_c$ is much smaller compared to the other terms, due to the typical sizes of ω and $\mathbf{T}_c$. Firstly, the magnitude of ω is normally much smaller than that of translation $\mathbf{v}$, unless the rotation

[1]The apparent difference in the operations done to the antipodal pair of optical flows (summation in (Lim and Barnes, 2007) and subtraction in (Thomas and Simoncelli, 1994)) is a consequence of the sign difference introduced by the cross product operation in (Thomas and Simoncelli, 1994); the eventual expressions in both formulations are essentially the same. In this connection, we would also like to observe that it is incorrect to say that the subtraction operation of (Thomas and Simoncelli, 1994) will face problem when the antipodal points are equally far away, as claimed by both sets of authors. The so-called angular translation flows of an antipodal pair in (Thomas and Simoncelli, 1994) are pointing in opposite directions; thus subtracting them would not make them vanish!

is very dominant in the system. Secondly, the radius r in $\mathbf{T}_c$ is usually much shorter than 1 m in both man-made and biological systems. Thus, multiplying all the terms on the RHS of Equation (3.9) by r further reduces their magnitude. As a consequence, the induced translation terms on the RHS are negligibly small compared to the true parallax terms on the left-hand-side. Due to the smallness of those terms, it would be numerically questionable to solve all the unknowns in Equation (3.9) directly, whether via a nonlinear method or via a linearizing scheme (by ignoring the dependency among the unknowns).

Viewed in another way, the quasi-parallax formulation are not suited for estimating both the translation and the rotation together, as the effect of the rotation is very weakly represented in the quasi-parallax via the induced translation. Trying to fit all the unknowns in one go would result in overfitting and produce a biased solution that is noise sensitive.

Instead, we find that numerically it is much more stable to first ignore the residual terms on the RHS of Equation (3.9) and solve the homogeneous system $\mathbf{A}\mathbf{x}_1 = 0$ via Total Least Squares (TLS). With the approximate translation estimate, we proceed to Stage 2 (Section 3.3.2.2) to solve for the rotation. We then substitute the rotation estimate $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ back into $\mathbf{x}_2$ of Equation (3.9) and form a new equation:

$$\begin{aligned} \underbrace{[\mathbf{A}, \mathbf{B}\mathbf{x}_2]} \cdot \underbrace{[\mathbf{x}_1, r]^T} &= 0 \\ \widetilde{\mathbf{A}}_{N\times 4} \cdot \quad \widetilde{\mathbf{x}}_{\mathbf{1}} &= 0 \end{aligned} \tag{3.11}$$

Assume the singular values of $\widetilde{\mathbf{A}}$ are $\tau_1 \geqslant \ldots \geqslant \tau_4 \geqslant 0$. If the computed $\mathbf{x}_2$ is accurate enough, or alternatively, the whole residue $r\|\mathbf{B}\mathbf{x}_2\|$ is insignificant compared to the other terms, τ_4 will be close to zero, since Equation (3.11) is homogenous. It follows that we can use this condition to check if we need to refine the current estimate of the translation.

As the absolute value of τ_4 is also affected by the level of noise, we consider instead both the value of τ_4 and the ratio $\rho = \frac{\tau_3}{\tau_4}$. If $\rho > 100$ and $\tau_4 < 0.3$, we deem the rotational residue $r\|\mathbf{B}\mathbf{x}_2\|$ insignificant or well estimated, so that Equation (3.11) can be satisfyingly regarded as a homogeneous system of equations[2]. In this case, we accept the current solution $(\hat{\mathbf{v}}, \hat{\omega})$ as correct. Otherwise, we recompute the translation by solving Equation (3.11) with TLS. With the new translation estimate, the rotation estimate can also be refined via Stage 2. In principle, we can iterate the above process until both the translation and rotation estimates converge, However, we find that under most cases tested in our simulation, $\widetilde{\mathbf{A}}$ is rank-deficient and there is no need to refine the estimates. Even in cases when the induced terms are not negligibly small (for instance, caused by dominant rotation), the solution converges after one iteration.

Based on the above observations, we propose the method in **Algorithm 1**. Note that the recovery does not require the knowledge of r and the recovered translation $\hat{\mathbf{v}}$ is in the form of $(sU,\ sV,\ sW)^T$, up to an unknown scale factor s which scales the magnitude of $\hat{\mathbf{v}}$ to unity.

3.3.2.2 Stage 2: Recover the global rotation

It is not advisable to compute (α, β, γ) from Equation (3.9), since it comprises chiefly of global translation (U, V, W). Instead, we revert to the epipolar constraint of each camera and use all the available feature points (not necessarily matching points) to recover rotation. Substituting the estimated $(sU,\ sV,\ sW)^T$ into the individual epipolar constraints

[2]Given a approximation of (α, β, γ), checking the rank condition of $\widetilde{\mathbf{A}}$ works better than checking the rank of $\mathbf{A}$ directly, especially when the induced translation terms might not be vanishingly small.

Algorithm 1 Linear Quasi-parallax Algorithm

1: Recover the translation estimate $\hat{\mathbf{v}}$ from $\mathbf{A}\mathbf{x}_1 = 0$, which is to be solved by TLS approach in Section 3.4.
2: Given $\hat{\mathbf{v}}$, compute $\hat{\omega}$ as described in **Algorithm 2**.
3: Check if current estimate ($\hat{\mathbf{v}}$,$\hat{\omega}$) obeys the rule of positive depth. If not, flip $\hat{\mathbf{v}}$ by 180° and go back to Step 2 to recompute rotation with $-\hat{\mathbf{v}}$.
4: Check if the induced translation is insignificant. If the following condition is satisfied, the algorithm regards the induced translation as insignificant: it stops and returns current $(\hat{\mathbf{v}}, \hat{\omega})$ as the global motion. Otherwise, the algorithm proceeds to Step **5**.
[Condition:] Substitute $\hat{\omega}$ into $\mathbf{x}_2$ to form $\widetilde{\mathbf{A}}$ in Equation (3.11). If $\rho > 100$ and $\tau_4 < 0.3$, we consider $\widetilde{\mathbf{A}}$ as having rank 3 and accept current solution $(\hat{\mathbf{v}}, \hat{\omega})$ as correct.
5: $\widetilde{\mathbf{A}}$ is full rank. Recompute translation by using TLS to solve Equation (3.11) for $\widetilde{\mathbf{x}}_\mathbf{1}$. Then $\hat{\mathbf{v}}$ is returned as the first three components of $\widetilde{\mathbf{x}}_\mathbf{1}$. Rotation is also refined using this updated $\widetilde{\mathbf{x}}_\mathbf{1}$. Repeat Steps 4 and 5 until convergence.

of the two cameras, we obtain a system of equations in the form of:

$$\mathbf{M} \cdot [\underbrace{\alpha, \beta, \gamma}_{\Theta_1}, \frac{r}{s}\alpha, \underbrace{\frac{r}{s}\beta, \frac{r}{s}\alpha\beta, \frac{r}{s}\alpha\gamma, \frac{r}{s}\gamma\beta, \frac{r}{s}(\alpha^2 - \beta^2)}_{\Theta_2 = \frac{r}{s}\Phi}]^T = \mathbf{d} \tag{3.12}$$

where $\mathbf{M}$ is a data matrix, $\mathbf{d}$ is a measurement vector, Θ_1 and Θ_2 are two unknown vectors.

Denote (u_1, v_1) as the optical flow at the feature point (x_1, y_1) in camera 1 and (u_2, v_2) as the flow at (x_2, y_2) in camera 2. Then $\mathbf{m}_{cam1}$ and $\mathbf{m}_{cam2}$, respectively the rows of $\mathbf{M}$ arising from the measurement at (x_1, y_1) in camera 1 and (x_2, y_2) in camera 2, are given by:

$$\begin{aligned}
\mathbf{m}_{cam1} = [&f^2U - fWx_1 - x_1y_1V + y_1^2U, (f^2V - fWy_1 - x_1y_1U + x_1^2V), \\
&Wy_1^2 - Ufx_1 - Vfy_1 + Wx_1^2, -u_1f, -v_1f, y_1^2 - x_1^2, y_1f, -x_1f, x_1y_1] \\
\mathbf{m}_{cam2} = [&x_2y_2V - f^2U + fWx_2 - y_2^2U, fWy_2 - f^2V + x_2y_2U - x_2^2V, \\
&Ufx_2 - Wy_2^2 + Vfy_2 - Wx_2^2, -u_2f, v_2f, y_2^2 - x_2^2, y_2, -x_2, x_2y_2]
\end{aligned} \tag{3.13}$$

Defined in a similar manner, $\mathbf{d}_{cam1}$ and $\mathbf{d}_{cam2}$ are given by:

$$\mathbf{d}_{cam1} = f(v_1U - u_1V) + W(u_1y_1 - v_1x_1)$$
$$\mathbf{d}_{cam2} = W(v_2x_2 - u_2y_2) - f(u_2V_2 + v_2U_2) \tag{3.14}$$

Due to the existence of six higher order terms in Θ_2, Equation (3.12) cannot be directly solved by linear techniques without compromising the recovery performance. Even solving the equation nonlinearly is fraught with the danger of overfitting, since these six terms of Θ_2 contribute to Equation (3.12) insignificantly compared to the first three terms of Θ_1. For our application scenario, with its typical values of r and s, and with the (x, y) values arising from small to moderately small FOVs, the contribution of the second-order terms (i.e. $\frac{r}{s}\alpha, \frac{r}{s}\beta$) is typically one to two orders of magnitude smaller compared to that of the first three terms, and the contribution of the remaining four third-order terms is yet another order of magnitude smaller. Incorporating all these terms will result in an overly complex model that captures noises in the data. As before, we adopt the strategy of reducing the dimension of the problem and drop the four third-order terms. The remaining terms give rise to this system of equations:

$$\left[\ \mathbf{m}_1, \ldots, \mathbf{m}_5\ \right] \cdot \left[\ \alpha, \beta, \gamma, \tfrac{r}{s}\alpha, \tfrac{r}{s}\beta\ \right]^T = \mathbf{d} \tag{3.15}$$

where $\mathbf{m}_i$ is the i^{th} column in $\mathbf{M}$. Further ignoring the dependency among the unknown variables, we can compute the initial estimate $\hat{\omega}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0)^T$ from Equation (3.15) by the linear least squares technique.

The recovered $\hat{\omega}_0$ is used for the refinement step. Substituting $\hat{\omega}_0$ into Φ in Θ_2, we have $\hat{\Phi}_0 = \Phi|_{\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0}$. Rearranging Equation (3.12) leads to a standard linear system:

$$\left[\ \mathbf{m}_1,\ \ \mathbf{m}_2,\ \ \mathbf{m}_3,\ \ [\mathbf{m}_4 \ldots \mathbf{m}_9]\hat{\Phi}_0\ \right] \cdot \left[\ \alpha,\ \ \beta,\ \ \gamma,\ \ \tfrac{r}{s}\ \right]^T = \mathbf{d} \tag{3.16}$$

Solve the above equation for an updated rotation estimate. If necessary, we substitute the newly obtained estimate back into Equation (3.16) and solve it again for a more refined solution. This process can be repeated until the solution converges. Numerical tests show that the estimate always converges onto a global solution after one or two iterations. This fast convergence can be attributed to the small magnitude of those higher order terms, in comparison to the first three terms of Θ_1 in Equation (3.12). **Algorithm 2** presents our linear rotation estimation algorithm, carried out in three steps.

Algorithm 2 Linear Rotation Recovery Algorithm

1: **Linearize:** Directly solve Equation (3.15) as a linear system, with the solution given by $\mathbf{M_1}^{+}\mathbf{d}$ where $\mathbf{M_1}^{+}$ is the pseudo-inverse of the matrix $\left[\ \mathbf{m}_1, \ldots, \mathbf{m}_5\ \right]$.
2: **Refine:** Given the estimate $\hat{\omega}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0)^T$, compute $\hat{\Phi}_0 = \Phi|_{\omega_0}$ to form Equation (3.16). Solve Equation (3.16) by linear least squares for a new estimate of ω.
3: **Iterate:** Iterate Step 2, if necessary, until convergence.

3.3.3 Extended Quasi-parallax for Multiple Camera Pairs

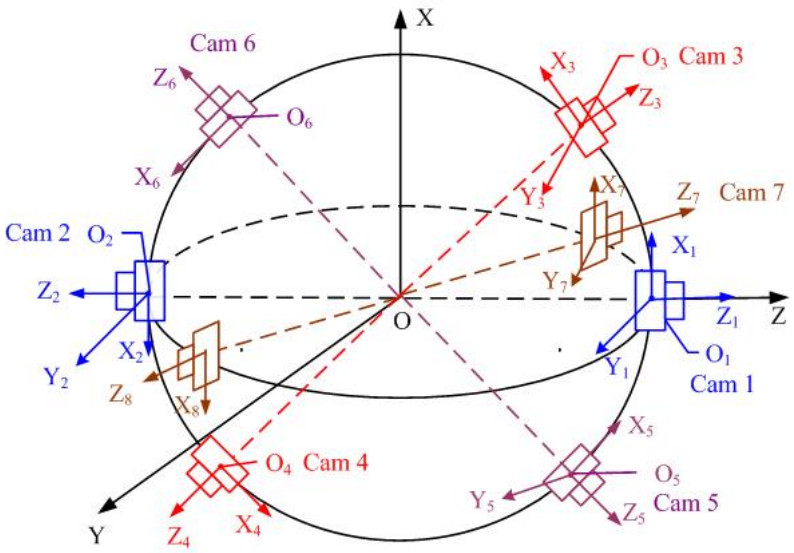


Figure 3.4: Multiple camera pairs configured as a compound eye. The global coordinate is $\{$O-XYZ$\}$ and the k^{th} camera's coordinate is $\{\mathrm{O_k} - \mathrm{X_k Y_k Z_k}\}$. The system is made up of camera pairs placed at diametrically opposite positions, such as the camera pair 3 and 4.

Figure 3.4 shows a system of multiple camera pairs configured in the manner of the insect compound eye. It consists of many small-FOV cameras, situated on the surface of a sphere. Each individual camera represents an ommatidium and has a visual axis which points outward in the direction of the surface normal of the sphere. Any two diametrically opposing cameras (e.g. Camera 3 and Camera 4) can be considered as a lateral eye pair discussed in the preceding sections, with a matrix equation similar to Equation (3.9). Suppose N is the total number of matching points collected from all the pair of cameras. Collecting the respective equations from all the pairs in the system, we have:

$$\mathbf{A}^*_{(N\times 3)}\mathbf{x}_1 = r\mathbf{B}^*_{(N\times 9)}\mathbf{x}_3 \tag{3.17}$$

where $\mathbf{x}_3 = [\alpha, \beta, \gamma,\ \alpha\beta,\ \beta\gamma,\ \alpha\gamma,\ \alpha^2,\ \beta^2,\ \gamma^2]^T$. Similar to $\mathbf{A}$ and $\mathbf{B}$ in Equation (3.9), the data matrices $\mathbf{A}^*$ and $\mathbf{B}^*$ depend on the optical flows and the matching point positions. In addition, $\mathbf{A}^*$ and $\mathbf{B}^*$ also contain terms decided by the orientation and

displacement of each camera pair with respect to the global coordinate.

Translation recovery is rather straightforward: we just need to substitute Equation (3.17) into **Algorithm 1** on page 38 and follow the procedure accordingly. Once the translation is recovered, we then proceed to the rotation estimation stage, described in **Algorithm 2** on page 40. The original system of equations in Equation (3.12) is slightly modified to:

$$\mathbf{M}^* \cdot \left[\underbrace{\alpha, \beta, \gamma,}_{\Theta_1} \underbrace{\frac{r}{s}[\alpha, \beta, \gamma, \alpha\beta, \alpha\gamma, \gamma\beta, \alpha^2, \beta^2, \gamma^2]}_{\Theta_3 = \frac{r}{s}\Phi^*} \right]^T = \mathbf{d}^* \tag{3.18}$$

where $\mathbf{M}^*$ is the data matrix, $\mathbf{d}^*$ is the measurement vector, and Θ_1 and Θ_3 are two unknown vectors. Compared to the original Equation (3.12), we now have 3 more higher order unknown terms. Such terms arise because in many lateral eye pairs, their optical axes are no longer aligned with the global Z-axis, introducing more coupling terms between the induced translation and the rotation.

The **Linearize** step in Equation (3.15) needs to be modified too to incorporate the $\frac{r}{s}\gamma$ term:

$$\left[\mathbf{m}_1^*, \ldots, \mathbf{m}_6^* \right] \cdot \left[\alpha, \beta, \gamma, \tfrac{r}{s}\alpha, \tfrac{r}{s}\beta, \tfrac{r}{s}\gamma \right]^T = \mathbf{d}^* \tag{3.19}$$

where $\mathbf{m}_i^*$ is the i^{th} column in $\mathbf{M}^*$. The rest of the procedure is the same.

3.4 TLS and Data Normalization

In real images, the data matrix $\mathbf{A}$ of a system of linear equations $\mathbf{A}\mathbf{x}_1 = 0$ is inevitably perturbed by noise. From a statistical perspective, the Total Least Squares (TLS) approach is better suited to deal with such *errors-in-variables* (EIV) models, as compared

to the classical Linear Least Squares techniques (Leedan and Meer, 2000). The TLS approach has a restrictive requirement that the error covariance matrix $\mathbf{C}$ associated with $\mathbf{A}$ should be an identity matrix scaled by an unknown scale factor. As this condition is violated in many applications as well as in our case, we need to perform data normalization on the matrix $\mathbf{A}$. We will demonstrate how this can be done in the case of Equation (3.10). The normalization matrix for $\tilde{\mathbf{A}}$ in Equation (3.11) could be obtained in a similar spirit.

The following assumptions are required:

1. The noises in the optical flows of both cameras are additive, i.i.d and Gaussian, with σ_v^2 as the noise variance.

2. The matching points' positions are not corrupted by noise.

With the preceding assumptions, the error in the i^{th} observed vector can be written as:

$$\mathbf{\Delta a}_i = [\Delta v_1^i + \Delta v_2^i, \Delta u_2^i - \Delta u_1^i,\ (\Delta u_1^i - \Delta u_2^i)\frac{y^i}{f} - (\Delta v_1^i + \Delta v_2^i)\frac{x^i}{f}] \tag{3.20}$$

where $(\Delta u_c^i, \Delta v_c^i)$ represents the Gaussian noise added to the flow at the i^{th} matching pair in the c^{th} camera. Denoting the covariance matrix associated with $\Delta \mathbf{a}_i$ as $\mathbf{C}_i$ ($i = 1 \ldots N$), we have:

$$\mathbf{C}_i = 2\sigma_v^2 \begin{bmatrix} 1 & 0 & -\frac{x^i}{f} \\ 0 & 1 & -\frac{y^i}{f} \\ -\frac{x^i}{f} & -\frac{y^i}{f} & \frac{x^{i2}+y^{i2}}{f^2} \end{bmatrix} \tag{3.21}$$

Clearly $\mathbf{C}_i$ is not the identity matrix as required. We should normalize the dataset to make the average covariance matrix closer to identity. The normalization scheme is carried out in the following order to obtain the normalization matrix $\mathbf{H}$.

1. Translate: Shift the centroid of the data to the image origin so that the off-diagonal terms in $\mathbf{C}_i$ are closer to 0.

2. Scale: Normalize the shifted matching points $(\frac{x^i}{f}, \frac{y^i}{f})$ so that they have a scatter closer to a unit circle. As a result, the last diagonal entry $\frac{x^{i2}+y^{i2}}{f^2}$ in $\mathbf{C}_i$ will be nearer to 1.

Normalized by $\mathbf{H}$, the homogeneous equation $\mathbf{A}\mathbf{x}_1 = 0$ becomes:

$$(\mathbf{A}\mathbf{H})(\mathbf{H}^{-1}\mathbf{x}_1) = 0 \quad \Rightarrow \quad \bar{\mathbf{A}}\bar{\mathbf{x}}_1 = 0 \tag{3.22}$$

The solution for $\bar{\mathbf{x}}_1$ is given by the eigenvector associated with the smallest singular value of $\bar{\mathbf{A}}$. The original translation $\mathbf{x}_1$ is then recovered as $\mathbf{H}\bar{\mathbf{x}}_1$.

After normalization, not only the error covariance matrix $\mathbf{C}$ becomes much closer to identity, the condition number of $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$ is also much smaller, which means that the system is more robust to noise perturbations.

3.5 Numerical Characterization

3.5.1 Extended Bundle Adjustment

In this section, we intend to compare our method against the gold standard solution obtained by the BA algorithm using Fisher Information Matrix. The purpose of the comparison is not intended to establish the superiority of our method over BA or otherwise; in any case, the BA method is usually applied to scenarios with longer baselines than the differential displacements being considered here. In our system, where the scene points cannot be tracked over a large number of views, the bundle of visual rays being adjusted in the BA are "local" to each camera (over its successive views), although in our formulation, the adjustment does obey the constraint that the individual camera motions must arise from the same global motion. When the field of view of each individual camera is small, difficulties might arise and it is not at all clear if BA would have a better perfor-

mance than our quasi-parallax formulation. The purpose of the following comparison is to shed some light on this issue.

The original BA deals with a single camera and thus needs to be extended to a system of two laterally placed cameras. The outline of the extended BA is given out in **Algorithm 3**. The extension of the preceding algorithm to the case of multiple camera pairs is straightforward and will be carried out in the next section for experimental comparison.

Algorithm 3 Extended Bundle Adjustment

1: **Initialize:** Using linear subspace method, solve camera motion $(\mathbf{v}_c, \omega_c)$ $(c = 1, 2)$ separately and compute initial world depths $\mathbf{Z}_0$ for all feature points. The initial estimates of the global motion $(\mathbf{v}_0, \omega_0)$ is given by $\mathbf{v}_0 = (\mathbf{v}_1 + \mathbf{R}\mathbf{v}_2)/2$, and $\omega_0 = (\omega_1 + \mathbf{R}\omega_2)/2$ in this simple configuration, where $\mathbf{R}$ is a diagonal matrix with $[-1, 1, -1]$ on the diagonal.

2: **Estimate:** For every feature point $\mathbf{p}_i$ $(i = 1 \ldots M_c)$ in the c_{th} camera, compute the back-projected flow $(\hat{u}_c^i, \hat{v}_c^i)$ using the global motion estimate.

3: **Iterate:** Minimize the following nonlinear cost function using Levenberg-Marquardt algorithm over $M_1 + M_2 + 6$ variables where M_1 and M_2 are the number of feature points in Camera 1 and Camera 2 respectively. There are $M_1 + M_2$ depth unknowns forming the vector $\mathbf{Z}$ and 6 global motion parameters. For the sake of simplicity, we assume $M_1 = M_2$ and r is known.

$$J(\mathbf{v}, \omega, \mathbf{Z}) = \sum_{c=1}^{2} \sum_{i=1}^{M_c} [(\hat{u}_c^i - u_c^i)^2 + (\hat{v}_c^i - v_c^i)^2] \tag{3.23}$$

3.5.2 Quasi-parallax versus Bundle-adjustment

We investigate the effectiveness of the quasi-parallax formulation in removing the inherent ambiguity between the translation and the rotation, and compare it against that of the gold standard BA formulation. Fisher information matrix is used to compute the covariance between motion parameters. Large off-diagonal entries in this matrix indicate an inherent ambiguity between the corresponding parameters. It is well-known that a single camera suffers from the ambiguity between the U and β pair, and the V and α pair.

According to Pless (Pless, 2004), if we assume a Gaussian distribution for the errors in the measured optical flows, the Fisher matrix of a multiple camera system is defined as:

$$\mathbf{F} = \sum_{\mathbf{k}\in D}(\sum_{i=1}^{N} \frac{\partial \mathbf{q}_i}{\partial \mathbf{k}}^T \frac{\partial \mathbf{q}_i}{\partial \mathbf{k}}) \mid_{\mathbf{k}=(\mathbf{v},\omega,Z_1 \ldots Z_N)} \tag{3.24}$$

where N is the total number of feature points and $\mathbf{q}_i$ is the optical flow measured at the i^{th} feature point. $\mathbf{k}$ is a vector of unknown parameters which contains the global motion and the depths of all the feature points. Each $\mathbf{k}$ in the parameter set D defines a motion-scene configuration and has an associated Fisher matrix. The numerical integration of this matrix over many samples from D, characterizes the behavior of a camera system in the environment described by D.

The camera and the motion-scene configuration for computing the Fisher matrices is as follows. Each camera has a 40° FOV and views a different scene with depths ranging from 3m to 7m in one camera, and from 5m to 10m in the other. As we will see later, this asymmetrical depth distributions for the two cameras is crucial to bring out the hidden ambiguity in the extended BA method. Both the translation and the rotation are sampled uniformly, with their respective norms equal to 1 and 0.01 respectively.

For a laterally-placed camera pair, the extended BA simply collects all the available flows from the two cameras. Thus the measurement vector $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N}$ is $[\mathbf{u}_1; \mathbf{v}_1; \mathbf{u}_2; \mathbf{v}_2]$, as no interaction exists between the two cameras at the level of optical flows. The overall Fisher matrix is equal to the addition of the individual Fisher matrices for each camera.

As shown in Table 3.1(b), the bas-relief ambiguity still looms large in the extended BA. This contradicts Pless' claim in (Pless, 2004) that no ambiguity exists in this lateral eye set-up. This discrepancy can be attributed to the fact that in his simulations, both cameras were viewing scenes with identical depth distributions. Under such setting,

Table 3.1: Fisher matrices of two methods with ambiguities highlighted, QP refers to Quasi Parallax and BA refers to Bundle adjustment

U	V	W	α	β	γ
1	0	0	0	0.002	0
0	1	0	-0.002	0	0
0	0	0.088	0	0	0
0	-0.002	0	0.001	0	0
0.002	0	0	0	0.001	0
0	0	0	0	0	0

(a) QP-based with FOV$= 50^\circ$ each

U	V	W	α	β	γ
0.03	0	0	0	**0.035**	0
0	0.03	0	**-0.036**	0	0
0	0	0.002	0	0	0
0	**-0.036**	0	0.994	-0.004	0
0.035	0	0	-0.004	1	0
0	0	0	0	0	0.079

(b) BA-based with FOV$= 50^\circ$ each

U	V	W	α	β	γ
1	0	0	0	0.001	0
0	1	0	-0.001	0	0
0	0	0.947	0	0	0
0	-0.001	0	0.001	0	0
0.001	0	0	0	0.001	0
0	0	0	0	0	0

(c) QP-based with FOV$= 100^\circ$ each

U	V	W	α	β	γ
0.017	0	0	0	0.003	0
0	0.017	0	-0.003	0	0
0	0	0.016	0	0	0
0	-0.003	0	0.981	0.005	0
0.003	0	0	0.005	1	0
0	0	0	0	0	0.352

(d) BA-based with FOV$= 100^\circ$ each

the corresponding ambiguities of the individual cameras, say the confusion between U and β, manifest in covariance terms with the same magnitude but different signs. Thus if added up, these ambiguities canceled each other and the entry of $U\beta$ in the overall Fisher matrix became zero. Hence the "zero ambiguity" phenomenon in (Pless, 2004) can be regarded as a case of special scene type; here, a more general, asymmetric depth scene reveals that BA still suffers from the bas-relief ambiguity in the case of small and moderate field of view.

In our quasi-parallax formulation, the input measurement is chosen to facilitate translation pickup. The Fisher matrix is thus based on the matching points, that is, $\mathbf{Q}$ is given by $[\mathbf{u}_1 - \mathbf{u}_2; \mathbf{v}_1 + \mathbf{v}_2]$. It can be seen from the large values of the first three diagonal entries of Table 3.1(a) that the translation parameters are picked up well using our formulation. The last three diagonal entries also revealed that this formulation is not suitable for rotation recovery and indeed we used a separate step in Section 3.3.2.2 to

recover the rotation. Note that since ω_z is not present in our $\mathbf{Q}$, all the associated entries are zero.

Tables 3.1(c) and (d) show the Fisher matrices of the same two methods with identical setup, except that the field of view of each camera is now $100°$. We can see that the bas-relief ambiguities are very much reduced, especially in the case of the extended BA. The results suggest that with visual field that covers the entire visual sphere, the bas-relief ambiguities do vanish. Thus, the fact that it is only doing "local" bundle adjustment might not matter in this case.

Chapter 4

Other Formulation of Linear Quasi-Parallax

Besides the lateral eye model discussed above, there are other biological eye systems that can leverage the linear quasi-parallax. The following sections will explore how our quasi-parallax can be adapted to 1) a new imaging sensor inspired by spider eyes and 2) a pair of eyes with parallel optical axis pointing in the same directions. These kind of eyes differ from those in the previous chapter in that the eyes face the same directions.

4.1 Aranead Eye

A common eye model which exists in various families of actively hunting spiders inspires us to propose a new imaging sensor termed as the Aranead Eye. These spiders catch prey without a web; thus their eye structure and the attendant visual acuity is of paramount importance. Despite slight variation in the eye arrangement, the general topology of these different families of spiders is largely the same. They all have two to

three rows of eyes, each with its own retina and lens, spanning a wide field of view. As in Fig 4.1, a jumping spider has eight eyes, each with its own retina and lens, arranged into two rows. The front row has four eyes comprising two forward-facing pairs. The second row has two smaller eyes on either side. Back on the side of the head is another pair of eyes. This structure not only gives the spider a large field of view, but also enables it to rapidly pick up the movement of the prey.

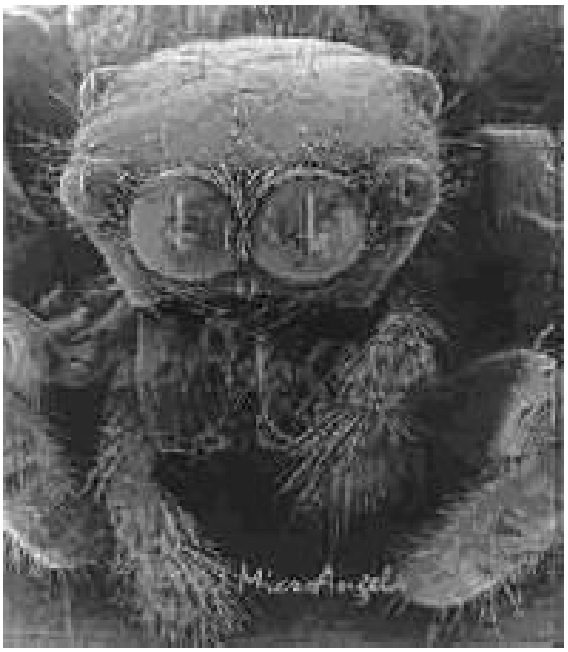


Figure 4.1: The eyes of a jumping spider

Such wide field of view can potentially disambiguate the coupling between rotation and translation for accurate motion recovery. Moreover, each individual eye is not unlike conventional pinhole camera, so that its resolution (especially the frontal eyes) is far better than that of compound eyes. This allows the spiders to detect prey quickly. Most importantly, SfM benefits a lot from the geometrical constraint afforded by the Aranead eye arrangement, namely, there exists at least two parallel rows of similarly oriented eyes.

4.1.1 System Set-up

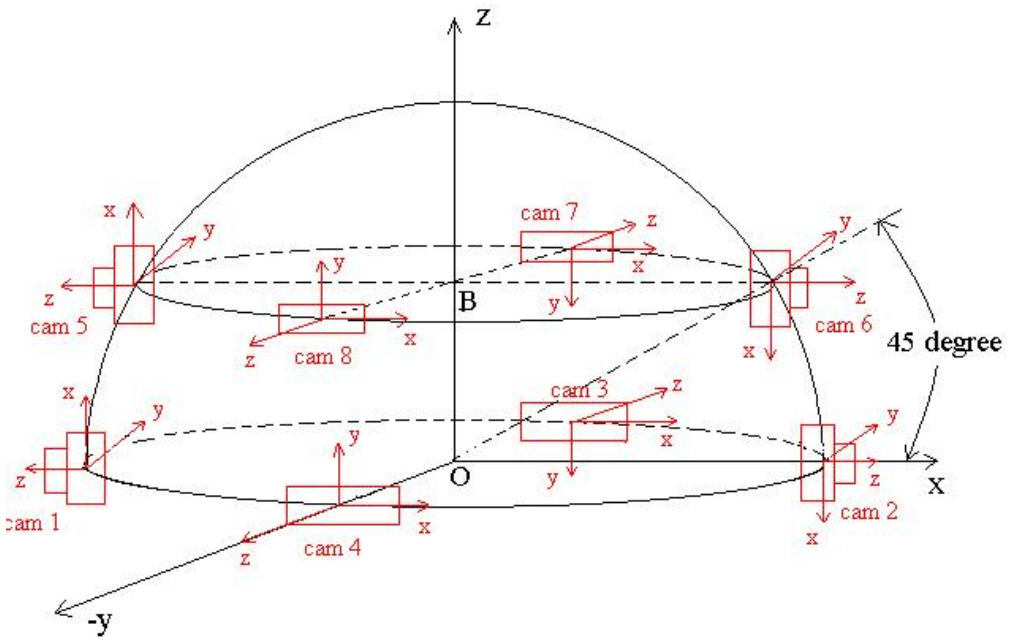


Figure 4.2: Configuration of the Aranead Eye

Figure.4.2 shows the configuration of the Aranead eye, where eight cameras with small field of view are mounted on the surface of a hemisphere. The hemisphere is built on a platform whose 3D motion is referred to as the global motion. The world coordinate's origin *O* is set at the center of the hemisphere. Four out of the eight cameras are on the equator with viewing directions emanating from the center. The remaining four cameras are positioned on the smaller circle whose center is *B*; they have the same orientations as that of the camera below them on the equator. We group these eight cameras into four pairs (camera *i* pairs with camera (*i*+4), *i*=1..4); thus each pair lie on a great circle and have the same viewing direction and orientation. All the optical centers of the eight camera are set on the surface of the hemisphere; the radius of the equator is *r* and the radius of the smaller circle is *t*.

4.1.2 Motion Recovery Algorithm

Denote the orientation and translation of camera *C* relative to the world reference frame as ***R*** and ***T*** respectively. Assume the hemisphere undergoes a global motion given by translation $\mathbf{v} = (U, V, W)^T$ and rotation ω=$(\alpha, \beta, \gamma)^T$. The individual 3D motions of some of the cameras induced by the global motion of the platform are listed below:

$$Camera\ 1: \quad \omega_1 = [\ \gamma \quad \beta \quad -\alpha\ ]^T \qquad \mathbf{v}_1 = [\ W + \beta r \quad v - \gamma r \quad -U\ ]^T \tag{4.1}$$

$$Camera\ 2: \quad \omega_2 = [\ -\gamma \quad \beta \quad \alpha\ ]^T \qquad \mathbf{v}_2 = [\ -W + \beta r \quad V + \gamma r \quad U\ ]^T \tag{4.2}$$

$$Camera\ 5: \quad \omega_5 = \begin{bmatrix} \gamma & \beta & -\alpha \end{bmatrix}^T \qquad \mathbf{v}_5 = \begin{bmatrix} W + \beta t & V - \alpha t - \gamma t & -U - \beta t \end{bmatrix}^T \tag{4.3}$$

$$Camera\ 6: \quad \omega_6 = \begin{bmatrix} -\gamma & \beta & \alpha \end{bmatrix}^T \qquad \mathbf{v}_6 = \begin{bmatrix} -W + \beta t & V - \alpha t + \gamma t & U + \beta t \end{bmatrix}^T \tag{4.4}$$

For general motion that is non-lateral, i.e. $W \neq 0$, we can eliminate W/Z from the optical flow equation and obtain the following equation:

$$(x - x_0)[v - \alpha(\frac{y^2}{f} + f) + \frac{\beta xy}{f} + \gamma x] = (y - y_0)[u - \frac{\alpha xy}{f} + \beta(\frac{x^2}{f} + f) - \gamma y)] \tag{4.5}$$

which can be re-written as:

$$[a_0 \quad a_1 \quad ... \quad a_{11}]^T \cdot [W \ U \ V \ \alpha W \ \beta W \ \gamma W \ U\alpha \ U\beta \ U\gamma \ V\alpha \ V\beta \ V\gamma] = 0 \tag{4.6}$$

where

$$\begin{aligned} &a_0 = -vx + uya_1 = fva_2 = -fua_3 = xfa_4 = yf \\ &a_5 = -(x^2 + y^2)a_6 = -(f^2 + y^2)a_7 = xya_8 = xf = a_3 \\ &a_9 = xy = a_7a_{10} = -(f^2 + x^2)a_{11} = yf = a_4 \end{aligned} \tag{4.7}$$

To differentiate between the coefficients $a_e, 1 \leq e \leq 11$ in Equation.4.6, we attach a subscript *i* after the coefficients of the i^{th} camera, e.g. $a_{0,i}$ is the a_0 in the i^{th} camera. Assume that for each pair of cameras, there are two corresponding feature points at the

same image position, called the matching points, in their respective image plane (not necessarily projected from the same world point). Thus for the pair of cameras *i* and *j*, $a_{e,i}$ equals $a_{e,j}$ for $3 \leq e \leq 11$. For brevity $a_{e,i}$ and $a_{e,j}$ will henceforth appear as a_i for $3 \leq e \leq 11$ in later discussion. However $a_{0,i}, a_{0,j}$,$a_{1,i}, a_{1,j}$, $a_{2,i}, a_{2,j}$ could still be different as they are dependent on the optical flow at the feature point.

Take the 2^{nd} pair of cameras, camera 2 and camera 6, as example. We substitute Equation.(4.2) and Equation.(4.4) into Equation.(4.6) and obtain Equation.(4.8) and Equation.(4.9) for camera 2 and camera 6 respectively.

$$\begin{aligned} &a_{0,2}U + a_{1,2}(-W + \beta r) + a_{2,2}(V + \gamma r) + a_3(-\gamma U) + a_4(\beta U) + a_5(\alpha U) \\ &+a_6(-W + \beta r)(-\gamma) + a_7(-W + \beta r)\beta + a_8(-W + \beta r)\alpha + a_9(V + \gamma r)(-\gamma) \\ &+a_{10}(V + \gamma r)\beta + a_{11}(V + \gamma r)\alpha = 0 \end{aligned} \tag{4.8}$$

$$\begin{aligned} &a_{0,6}(U + \beta t) + a_{1,6}(-W + \beta t) + a_{2,6}(V - \alpha t + \gamma t) + a_3(-\gamma)(U + \beta t) + a_4\beta(U + \beta t) \\ &+a_5\alpha(U + \beta t) + a_6(-W + \beta t)(-\gamma) + a_7(-W + \beta t)\beta + a_8(-W + \beta t)\alpha \\ &+a_9(V - \alpha t + \gamma t)(-\gamma) + a_{10}(V - \alpha t + \gamma t)\beta + a_{11}(V - \alpha t + \gamma t)\alpha = 0 \end{aligned} \tag{4.9}$$

Subtracting Equation.(4.9) from Equation.(4.8), we get:

$$\begin{aligned} &(a_{2,6}t)\alpha + (a_{1,2}r - a_{1,6}t - a_{0,6}t)\beta + (a_{2,2}r - a_{2,6}t)\gamma + (a_{0,2} - a_{0,6})U + (a_{2,2} - a_{2,6})V \\ &+(a_{1,6} - a_{1,2})W + (a_8(r - t) + (a_{10} - a_5)t)\alpha\beta + (a_3t + (a_{10} - a_6)(r - t))\beta\gamma \\ &+(a_{11}(r - t) - a_9t)\alpha\gamma + (a_{11}t)\alpha^2 + (a_7(r - t) - a_4t)\beta^2 - a_9(r - t)\gamma^2 = 0 \end{aligned} \tag{4.10}$$

Similarly, subtracting camera 5's equation from camera 1's yields:

$$\begin{aligned} &(a_{2,5}t)\alpha + (a_{1,1}r - a_{1,5}t + a_{0,5}t)\beta + (a_{2,5}t - a_{2,1}r)\gamma + (a_{0,5} - a_{0,1})U + (a_{2,1} - a_{2,5})V \\ &+(a_{1,1} - a_{1,5})W + ((a_{10} - a_5)t - a_8(r - t))\alpha\beta + (a_3t + (a_6 - a_{10})(r - t))\beta\gamma \\ &+(a_{11}(r - t) + a_9t)\alpha\gamma - (a_{11}t)\alpha^2 + (a_7(r - t) + a_4t)\beta^2 + a_9(t - r)\gamma^2 = 0 \end{aligned} \tag{4.11}$$

It is straightforward to derive the other equations by subtracting camera 7 from camera 3 and camera 8 from camera 4 respectively. As a consequence of combining equations from strategically located pairs of cameras, only six rotational nonlinear terms, $\alpha\beta$, $\beta\gamma$, $\alpha\gamma$, α^2, β^2, γ^2, are present in Equation.(4.10) and Equation.(4.11) in addition to the six global motion parameters *U*, *V*, *W*, α, β, γ (similar equations arise for the remaining 2 pairs of cameras). This formulation is almost identical to Equation 3.9, except the composition of the rotational nonlinear terms. Thus we can employ the algorithm in **Algorithm 1** to solve for the global motion parameters with $\mathbf{x}_1 = [U, V, W]^T$, $\mathbf{x}_2 = [\alpha, \beta, \gamma, \alpha\beta, \beta\gamma, \alpha\gamma, \beta^2 - \alpha^2]^T$.

4.2 Parallel Camera Array

Most predatory animals, from spiders and crustaceans through birds to mammals and humans, tend to have the eyes facing in the same direction. This arrangement is, presumably, to take advantage of stereoscopic depth perception, which has been demonstrated in humans, monkeys, cats, and falcons, and is likely to be widespread across other species. It is unclear, however, whether the binocular field is used to take advantage of stereoscopic depth cues or merely for the improvement in image reliability by binocular comparison. As discussed in the previous chapter, many birds seem to forego maximising binocularity, presumably because stereopsis involves considerable neural processing and is too slow to control the estimation of distance and depth when a bird is landing upon a perch (Davies and Green, 1994). Instead, it is likely that the general function of binocularity is concerned with optic flow-fields (Martin and Katzir, 1999). In this chapter, we explore how having a binocular pair of eyes can aid motion analysis

without necessarily involving stereopsis and having to deal with the attendant problem of feature correspondence.

4.2.1 Literature review

We focus our literature review on the so-called stereo-motion integration works which involve a pair of cameras moving in time. The integration can be extended to involve multiple cameras, but this case has been reviewed in the earlier chapter and is thus omitted here. We also want to distinguish between the standard temporal stereo + motion problem, and the more restricted problem of just estimating disparity and 2D motion from two consecutive frames in a stereo sequence. Our review will be primarily limited to the former.

Some earlier works of motion-stereo integration are merely juxtaposition of the results from independent processing of the motion and stereo information (Kriegman, Triendl, and Binford, 1989; Symosek et al., 1990; Ayache and Faugeras, 1989; Grosso, Sandini, and Tistarelli, 1989). The final estimates of the 3-D structure are based on some combination of the outputs of these separate processes. These works can be classified under the modular integration scheme whereby the degree of interaction between the processing of individual sources of 3-D structure is loose.

In contrast to the modular approach, the processing of one type of visual information may depend on the presence of another type in a non-modular manner. In the case of perceived 3-D structure from the combination of binocular stereopsis and kinetic depth, there has been growing evidence that a modular approach fails to adequately describe human performance (Bradshaw and Rogers, 1993; Johnston, Cumming, and Lany, 1994; Nawrot and ., 1993; Tittle and Braunstein, 1993). On the basis of these results, it seems

clear that some degree of interaction exists between the processing of 3-D structure from motion and binocular stereopsis.

Among these non-modular approaches, some of them depend on the stereo module as the main source of depth information which is then used to guide the motion analysis.

These works depend on a single (and possibly erroneous) source for the depth. There are also approaches that are based upon more integrated relations of motion and stereo (Argyros and Orphanoudakis, 1997; P. and M.A., 1991; Grosso and Tistarelli, 1995; Li and Duncan, 1993; Shi, Shu, and Pan, 1994; Waxman and Duncan, 1986). In (P. and M.A., 1991; Li and Duncan, 1993; Shi, Shu, and Pan, 1994; Waxman and Duncan, 1986), the temporal derivative of disparity (equivalently, the difference of flow between the stereo pair) is exploited. The works by (P. and M.A., 1991; Shi, Shu, and Pan, 1994; Waxman and Duncan, 1986) are close to ours in the sense that they take the difference of the image velocities in the pair of binocular images and the constraint is valid for general motion. They derive a special relation between the ratio of the relative optic flow in a binocular pair of images and the disparity. Argyros et al (Argyros and Orphanoudakis, 1997) consider a moving pair of fixating camera with unknown stereo parameters. The binocular and motion cues are combined into a single equation which is solved by linearizing the coupled stereo and motion parameters. The system in (Li and Duncan, 1993)is limited to the case where the stereo setup only executes translational motion, and as a result, it does not require point to point correspondence (instead, gross correspondence of regions is needed). (Grosso and Tistarelli, 1995) extracts time-to-contact information from the temporal evolution of the stereo disparity.

More recently, works like (Stein and Shashua, 1998; Zhang, 1995; Zhang and Negahdaripour, 2008) exploit the constraint that governs the stereo disparity and the optical

flow given two pairs of stereo views. Other approaches like (Li and Sclaroff, 2008; Tao, Sawhney, and Kumar, 2001; Vedula et al., 2005) recovered 3d information from video in the form of 3d scene flow constraint. Typically, the 3d scene flow is solved from 2d optical flow and correspondences between N cameras ($N \geq 2$).

With the advent of more sophisticated techniques, we also see approaches using PDE, variational and factorization techniques, often dealing with long video sequences. (Strecha and Gool, 2002) weighs both the stereo and the motion correspondences at every iteration and the depths are solved through the evolution of a system of coupled, non-linear diffusion equations. Ho et al. (Ho and Chung, 2000) cast the problem in the factorization framework, whereby the stereo correspondence becomes one ordering the columns of the extended measurement matrix. The recovery of 3d scene flow has also been cast in a variational framework (Huguet and Devernay, 2007; Pons, Keriven, and Faugeras, 2007; Williams, Isard, and MacCormick, 2005). In one form or other, stereo disparity is still estimated.

Our work also fuses the stereo and motion cues right at the early stage and thus it too exhibits tight cue integration. What sets our work apart from all the above is that we do not perform stereo correspondence, either explicitly or implicitly. Instead, we compute the difference in optical flow between the stereo pair at the same image location (x, y). The resulting quantity is akin to parallax and facilitates translation recovery, yet we do not have the problem faced by the traditional approaches of getting parallax (where too few image points provide good parallax). In other word, our way of exploiting the information in the stereo pair does not just offer statistical advantage, but it geometrically disambiguates the coupling between translation and rotation. Furthermore, we do not need strict parallel stereo rig. Our formulation allows small deviation from the parallel

configuration. It also allows gaze in any direction as long as the stereo pair is approximately parallel.

4.2.2 Motion Recovery Algorithm

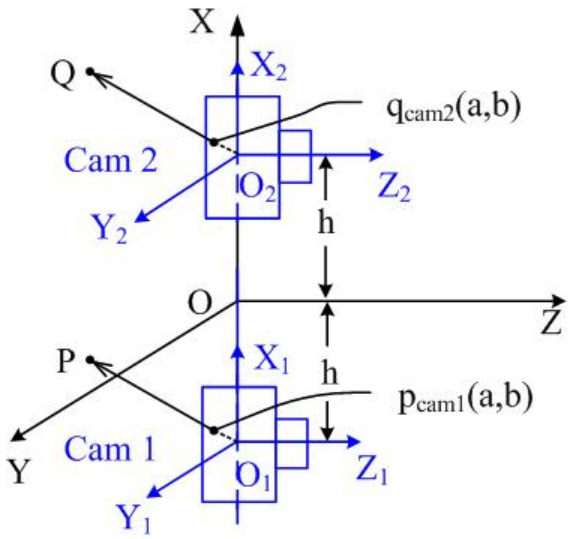


Figure 4.3: Set-up of the parallel camera array. $\mathbf{p}_{cam1}$ and $\mathbf{q}_{cam2}$ form a pair of matching points.

Figure 4.3 shows the set-up of our system. This setup is in a way a special case of those in Section 4.1, except that the global coordinate system is placed at the center and that $t = r$. The two cameras have parallel viewing directions and their optical centers are of equal distance (termed as h) from the origin O. The two cameras move together according to a global rigid motion expressed in the world coordinate system {O-XYZ}. The global motion is given by translation $\mathbf{v} = (U, V, W)^T$ and rotation ω=$(\alpha, \beta, \gamma)^T$.

A quasi-parallax motion field is obtained from pairs of matching points which experience identical rotational flows. At a matching point $\mathbf{p}$, we substitute the respective camera motions into the epipolar constraints and subtract them:

$$\mathbf{p}^T(\tilde{\mathbf{v}}_1\mathbf{r}_1 - \tilde{\mathbf{v}}_2\mathbf{r}_2) = -2\mathbf{p}^T(\omega \times \mathbf{T}) \times (\tilde{\omega}\mathbf{p}) \tag{4.12}$$

Normally we have multiple pairs of matching points in this form and collecting all

the N equations from the entire set gives us:

$$\mathbf{A}_{(N\times 5)}\mathbf{x}_1 = -\mathbf{B}_{(N\times 4)}\mathbf{x}_2 \tag{4.13}$$

$\mathbf{x}_1 = [U, V, W, h\beta, h\gamma]^T$, $\mathbf{x}_2 = h[\alpha\beta, \beta\gamma, \alpha\gamma, \beta^2 - \gamma^2]^T$. The i^{th} row of $\mathbf{A}$ is $[\ v_1 - v_2 \quad u_2 - u_1 \quad \frac{(v_2-v_1)x+(u_1-u_2)y}{f} \quad \frac{(u_1+u_2)y-(v_2+v_1)x}{f} \quad u_1 + u_2\ ]$, and the i^{th} row of $\mathbf{B}$ is $2[\ -x \quad \frac{y^2}{f} - f \quad \frac{xy}{f} \quad -y\ ]$. The form of Eq.(4.13)is also identical to Equation 3.9, thus **Algorithm 1** can be applied to solve this equation. This formulation is motivated by the theory that though some animals have frontally placed binocular eye pair, they are more used for motion computation rather than for stereo. One example is the bird, unless its eyes are manipulating an object in its grasp where object depths are very close.

The basic formulation for a pair of parallel cameras can be extended easily to the general case of an array of parallel camera pairs: the optical axis of the cameras are parallel but they are glancing sideway. That is, the z-axis are parallel but not pointing to the front so the x-axis are not aligned, as shown in Fig 4.4.

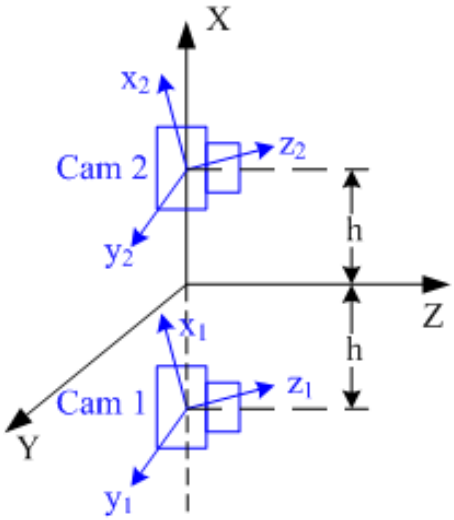


Figure 4.4: Set-up of a parallel camera pair where the gaze is sideway

Denote the world coordinates as $O - XYZ$ and the global motion is given by translation $\mathbf{v}_0 = (U, V, W)^T$ and rotation ω_0=$(\alpha, \beta, \gamma)^T$. The individual 3D motions of cam-

era 1 and camera 2 are $(\mathbf{v}_c, \omega_c), c = 1, 2$ and they are induced by the global motion:

$$\omega_1 = \omega_2 = R^T \cdot \omega_0 v_1 = R^T \cdot (v_0 + \omega_0 \times T) V_2 = R^T \cdot (v_0 - \omega_0 \times T) \tag{4.14}$$

where (R, T) is the orientation and translation of camera 1 relative to the world coordinate respectively, note that camera 2 has the same orientation as camera 1.

In this set-up, a quasi-parallax motion field similar to Equation 4.12 can be obtained by subtracting the respective epipolar constraints of camera 1 and camera 2 at the matching point:

$$\mathbf{p}^T(\tilde{\mathbf{v}}_1 \mathbf{r}_1 - \tilde{\mathbf{v}}_2 \mathbf{r}_2) = -2\mathbf{p}^T(R \cdot \omega \times \mathbf{T}) \times (\tilde{r} \cdot \omega \mathbf{p}) \tag{4.15}$$

Collect all the equations of matching points pairs and we have:

$$\mathbf{A}_{(N\times 5)} \mathbf{x}_1 = -\mathbf{B}_{(N\times 4)} \mathbf{x}_2 \tag{4.16}$$

$\mathbf{x}_1 = [U, V, W, h\beta, h\gamma]^T$, $\mathbf{x}_2 = h[\alpha\beta, \beta\gamma, \alpha\gamma, \beta^2 - \gamma^2]^T$. The form of Eq.(4.16)is also identical to Eq.(4.13), only the coefficients change.

The i^{th} row of $\mathbf{A}$ is $[(v_2 x - v_1 x + u_1 y - u_2 y)\sin\varphi + f(v_1 - v_2)\cos\varphi,\ \ f(u_2 - u_1)$

$f(v_2 - v_1)\sin\varphi + (v_2 - v_1)x\cos\varphi + (u_1 - u_2)y\cos\varphi,$

$(u_2 y - v_2 x + u_1 y - v_1 x)\cos\varphi - f(v_2 + v_1)\sin\varphi,\ \ f(u_1 + u_2)]$

The i^{th} row of $\mathbf{B}$ is $[\frac{1}{2}\sin 2\varphi(f^2 - x^2) + xf\cos^2 2\varphi,\ \ x^2 - y^2 + (f^2 - x^2)\cos^2\varphi - xf\sin 2\varphi.$

$-xy\cos\varphi - fy\sin\varphi,\ \ y(f\cos\varphi - x\sin\varphi)]$

Chapter 5

Experiments on Motion Recovery

Several experiments were conducted to test our quasi-linear parallax method under different scenarios and in different camera models. First we evaluate the method in a lateral eye system and then in one parallel eye with a configuration which can be regarded as a special case of Aranead eye. The motion recovery performance of our linear method was further compared against that of the extended BA. The evaluation criteria consist of the errors in direction (the angle between the estimated and the true motion in degree) for both the translation and rotation estimates, and the error in magnitude for the rotation estimate (the norm of the difference between the estimated and the true rotation). Note that the translation magnitude is not recoverable in our method since r is not known. Both methods were tested with varying translation-to-rotation ratio ε, computed as the ratio of the total magnitude of the translational flow and that of the rotational flow from all feature points. The image resolution used in this thesis is 512×512 pixels.

5.1 Lateral Eye

5.1.1 Experiment on Range Image for Lateral Eye

This set of experiment uses the Brown range image database (Lee and Huang, 2000) which contains many static natural scenes. Figure 5.1 shows the forest scene we used with depths ranging from 3m to 10m.

Figure 5.1: Range image of a forest scene used. Intensity represents depth with distant object looking brighter. Regions with no range data appears black. The image plane of camera 1 is denoted by a white rectangle

We endowed the scene with 3D motions, and projected the 3D scene points and their flows onto each camera's image plane. The resulting image points were matched across cameras. This scheme allows us to experiment with realistic scenes with its sparse and clustered feature distribution, and yet able to control the exact amount of noise added to the image. The noise added was a zero-mean Gaussian noise, isotropic in direction and with standard deviation equal to the Noise-to-Signal Ratio (NSR) times the average flow speed. We evaluated both the BA and our method under three motion types of ε =0.2, 1, 5 and subject to different levels of noise. The scale units for each type of motion were listed as follows.

Each simulation consisted of 300 trials for the linear method and 50 trials for the non-linear one. The error of each simulation was computed as the average of the errors from all the trials.

Table 5.1: In the forest scene, three types of motions corresponding to ε =0.2, ε=1, ε=5 are executed.

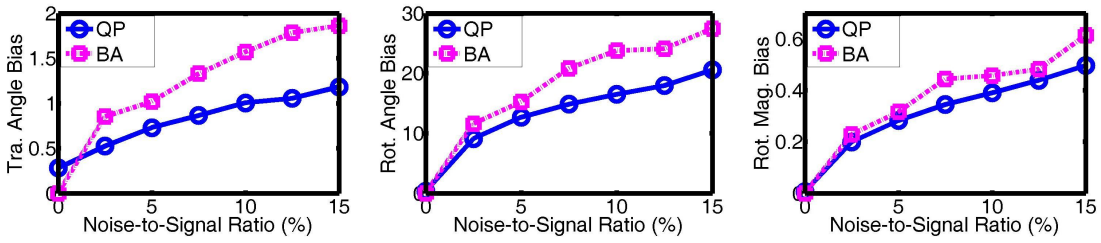
ratio ε	Motion Type	Translation: cm/s	Rotation: rad/s
ε=1	Balanced motion	[1, 3, 2]	[0.004, 0.003, 0.002]
ε=5	Dominant translation	[6, 12, 1]	[0.004, 0.003, 0.002]
ε=0.2	Dominant rotation	[1, 3, 2]	[0.01, 0.02, 0.016]

5.1.1.1 A lateral pair of cameras with narrow FOV

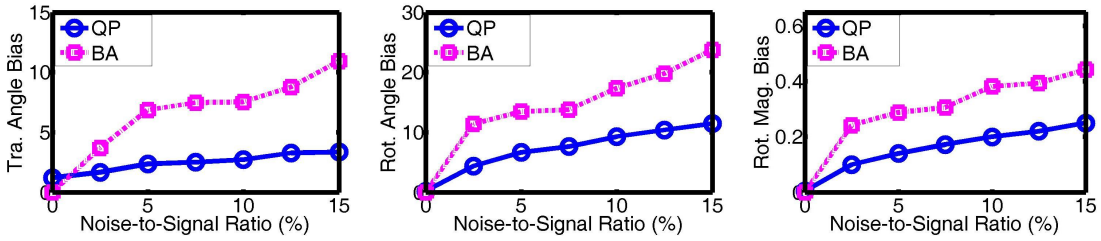
Our first experiment consists of a lateral pair of cameras with narrow FOV of 15° and $r = 0.1$ m. With such small FOV, bas-relief ambiguity is expected to be very severe and result in large errors in the solution. A total of 86 matching points were found for the linear method and 452 feature points for BA. In general, there are much fewer feature points for our method due to the need to satisfy the matching requirement[1].

Under all conditions, our linear method significantly outperformed the nonlinear BA in all aspects. Even at the highest noise level of 15%, the linear method was still robust. Take the recovery of the translation direction as an example (the left column in Figure 5.2). When ε=0.2, the estimation error of our method is only 30% of that of the nonlinear one. It is worth emphasizing that the linear method ran at a much less computational cost and required only about $\frac{1}{15}$ the computation time of the nonlinear method using Matlab on a 1.86GHz Pentium PC. Our superior performance can be attributed to the better resolving of the bas-relief ambiguity and the stable numerical behavior of TLS with proper data normalization. In contrast, the BA suffered from the ambiguity to a significant extent at such a narrow FOV and thus yielded strongly biased estimates. This confirms the earlier results obtained with the Fisher information matrix—even with a pair of laterally placed cameras, ambiguities may still exist if each camera only has narrow

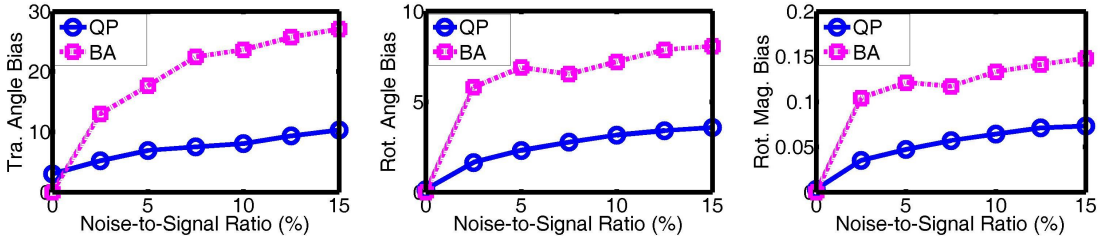
[1] If the distance between two feature points is no more than 5 pixels, they are considered as a matching pair.



(a) Estimation error when translation is dominant with ratio ε=5



(b) Estimation error when trans. and rot. are balanced with ratio ε=1



(c) Estimation error when rotation is dominant with ratio ε=0.2

Figure 5.2: Motion recovery of the two methods with a pair of cameras with narrow FOV of 15° each. QP stands for our Quasi-parallax method, while BA denotes bundle adjustment. The three columns, starting from the left, depict the estimation errors in translation direction, rotation direction, and rotation magnitude respectively. Note that the y-axis for different diagrams may have different scales.

FOV. We should also mention that the nonlinear BA method sometimes got trapped in local minima. If the error in direction of the estimate is greater than 30°, we deem that this has happened and ignore the trial result, in order not to unduly influence the results of the BA method.

With dominant translation in Figure 5.2(a), both methods recovered the translation accurately whereas the rotation estimates became worse off. Nevertheless, rotation recovery was much better in the linear method than the nonlinear BA. This can be attributed to the post-translation step used to estimate the rotation. Firstly, rotation estimation benefited substantially from a good translation estimate. Secondly, we are using

different aspects of the plenoptic function suited for rotation recovery. Thus the potentially unfavorable condition for rotation recovery caused by the dominant translation was ameliorated by our method. Decreasing ε expectedly improved rotation recovery while impairing translation recovery, as observed in both methods. However the nonlinear method was affected more adversely.

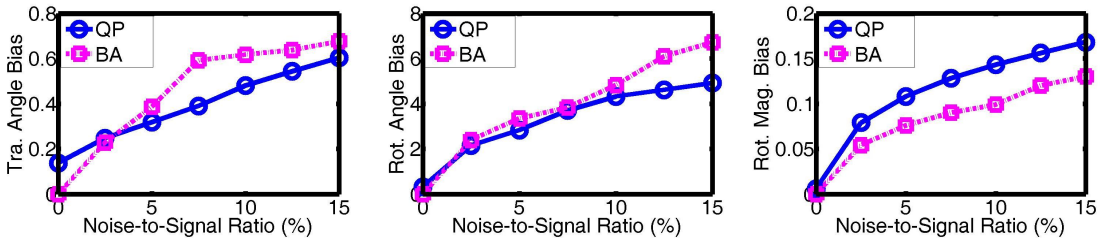
In terms of the translation direction recovery, when ε was reduced from 5 to 0.2, our linear method still managed to produce an acceptable estimate. In the case of 15% noise, our translation estimate was worsened by 9°. This is in contrast to the BA where the recovery deteriorated rapidly: under the same setting, the estimate error increased by 25°. This comparison indicated that under narrow FOV, the removal of rotational flow in the quasi-parallax formulation was very helpful for translation recovery.

The different degree of bas-relief ambiguity that still exists in both methods under narrow FOV can be illustrated from another perspective. Looking at the rotation recovery under increasingly dominant rotational flow, our method generated a significantly improved estimate. For example, when ε dropped from 5 to 1, the improvement of our rotation direction estimate was by 10° at 15% noise, whereas BA only improved slightly by 3°. The reason that no substantial improvement is seen in BA is precisely due to the bas-relief ambiguity that still exists under such narrow FOV, exacerbated by the high level of noise.

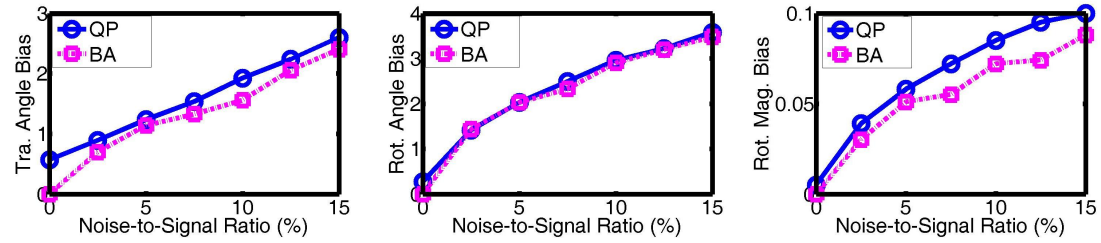
In sum, the better performance of the linear method under narrow FOV is due to the separate recovery of the translation and rotation, each using different aspects of the flow field most suited for their respective recoveries.

5.1.1.2 A lateral pair of vertebrate eyes

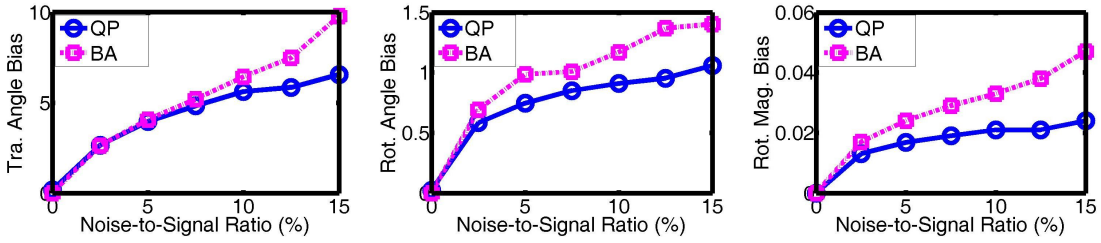
To mimic the two laterally-positioned eyes of some birds and vertebrates, we configure a visual system consisting of two diametrically opposite cameras with 50° FOV each and $r = 0.1$m. The results in Figure 5.3 show that the wide FOV of 50° has largely resolved the bas-relief ambiguity. Both methods performed better compared to that of the preceding experiment under narrow FOV. In particular, as the BA method no longer suffers from the bas-relief ambiguity, its performance became comparable and even outperformed ours in some instances, though overall, there is no clear winner among the two methods.



(a) Estimation error when translation is dominant with ratio ε=5



(b) Estimation error when trans. and rot. are balanced with ratio ε=1



(c) Estimation error when rotation is dominant with ratio ε=0.2

Figure 5.3: Motion recovery of the two methods with a pair of laterally-positioned cameras with 50° FOV each.

We picked 341 matching points for the linear method and a comparable number of

1058 feature points for BA. Under $\varepsilon = 1$ in Figure 5.3, the linear and the BA methods produced similar results, with the former's performance being slightly inferior. The maximal angular difference between the two was 0.2°, occurring in the translation recovery.

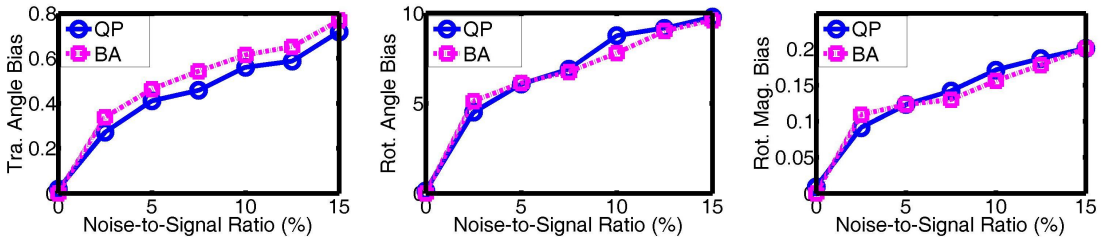
With a dominant translation of $\varepsilon = 5$, the quasi-parallax method actually performed better than the BA method in two aspects: both in the translation and rotation direction recovery. In comparison, the estimation of the rotation magnitude was not as good as the BA's.

As for the case of dominant rotation with $\varepsilon = 0.2$, the linear method outperformed BA significantly, especially in the translation estimation. With 15% noise, our estimate error is 4.1° less than BA's. This is not surprising if we recall that the undesirable influence of the strong rotational flow was removed from the estimation and thus its deleterious effect on the translation recovery was kept to a minimum.

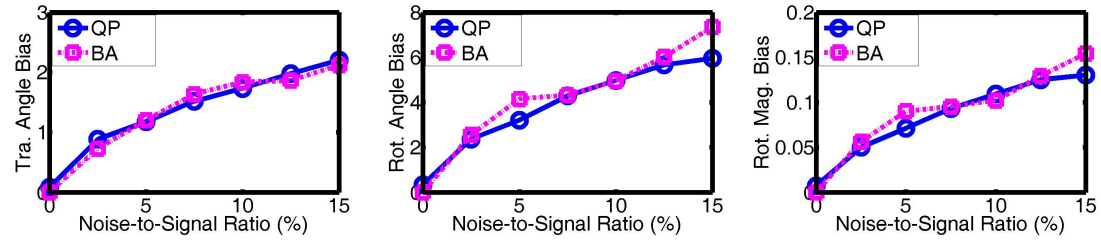
5.1.1.3 A compound eye with small number of facets

This section investigates the performance of both methods in a compound eye set-up. A multiple camera system with configuration similar to that of Figure 3.4 is employed, with the visual field of individual camera being 5° (which is the FOV of an individual ommatidium in the honeybee). In our simulation, each eye is made up of nine cameras, with all cameras arranged on the surface of a sphere with radius $r = 0.05$m. The linear method had 352 matching points, while the BA had 1279 feature points.

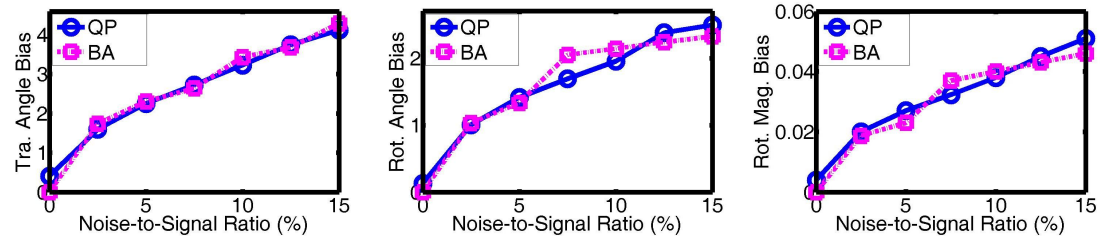
Under all conditions tested in Figure 5.4, the linear method and the BA produced almost identical results, given the wide coverage of the visual field. While our linear method achieved comparable accuracy to that of the nonlinear optimization techniques,



(a) Estimation error when translation is dominant with ratio ε=5



(b) Estimation error when trans. and rot. are balanced with ratio ε=1



(c) Estimation error when rotation is dominant with ratio ε=0.2

Figure 5.4: Motion recovery of the two methods in a compound eye with small number of facets.

it required much less computational cost, which is crucial for an agent (such as insect) with little computational resources and yet with a need to perform rapid visuomotor tasks.

5.1.2 Effect of Rotation-induced terms for Lateral Eye

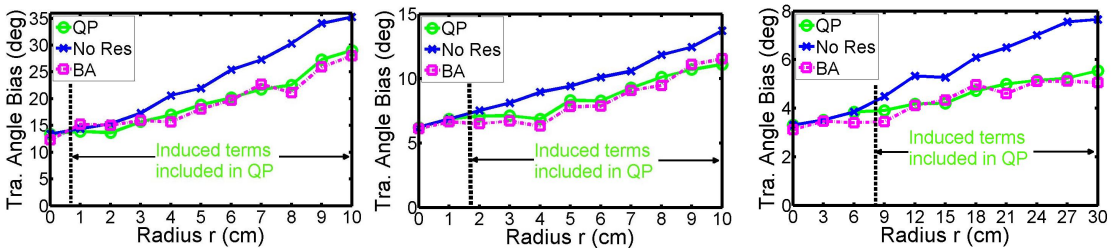
Here, we carried out an additional experiment to demonstrate the importance of accounting for the induced terms caused by the separation between each camera pair. We used the same compound eye setup as in the preceding experiment. The compound eye was viewing an indoor scene with depths ranging from 0.5m to 3m. A total of 185 matching points were selected for the linear method while 536 feature points were for bundle adjustment. The matched points of linear method were derived from the features points of

bundle adjustment, that's why they are much fewer. The flow input of the system was subject to a 10% isotropic noise.

Recall that our formulation modeled these induced terms as $r\mathbf{Bx_2}$ on the right hand side of Equation (3.9) and iteratively refined their estimates until Equation (3.11) is sufficiently close to a homogeneous system of equations. The performance of our method is here compared to one which sets r to zero, thereby ignoring the induced terms totally, and is effectively reduced to the antipodal constraint used in the spherical systems of (Lim and Barnes, 2007; Lim and Barnes, 2008; Thomas and Simoncelli, 1994). We compare the two schemes under three types of global motions: a rotation-dominant motion with ratio $\epsilon = 0.1$, a motion with significant rotation at $\epsilon = 0.3$, and finally, a balanced motion with $\epsilon = 1$. The details of each type of motion are tabulated as follows.

Table 5.2: In the indoor scene, three types of motions corresponding to ε =0.1, ε=0.3, ε=1 are executed.

ratio ε	Motion Type	Translation: cm/s	Rotation: rad/s
ε=0.1	Dominant rotation	[0.5, 1.5, 1]	[0.06, 0.12, 0.096]
ε=0.3	Significant rotation	[0.5, 1.5, 1]	[0.02, 0.04, 0.032]
ε=1	Balanced motion	[0.5, 1.5, 1]	[0.01, 0.0075, 0.005]



(a) Error with ratio ε=0.1 (b) Error with ε=0.3 (c) Error with ε=1

Figure 5.5: Errors in translation recovery when the induced terms are not modeled, with (a) $\epsilon = 0.1$, (b) $\epsilon = 0.3$ and (c) $\epsilon = 1$ respectively. "QP" and "BA" are as defined before, whereas the curve "No Res" refers to the case of ignoring the induced terms $r\mathbf{Bx_2}$ and solving the linear equation $\mathbf{Ax_1} = 0$ from Equation (3.9). The vertical dashed lines indicate the radius r above which our algorithm deems the induced terms as significant.

Referring to Figure 5.5, the results of the comparison show that given a moderately large r, the translation recovery in all types of motions were significantly improved if we explicitly modeled the induced terms in the estimation process. This improvement is especially obvious when the rotational flow was significant. For instance, with ratio $\epsilon = 0.3$ and $r = 0.1$m, the gap between the error in QP and that of ignoring the induced terms was 3.1°. When the rotation was dominant with ratio $\epsilon = 0.1$, modeling the induced terms at $r = 0.1$m improved the estimate by a large 6.4°. We also observe that even in the case of balanced motion ($\epsilon = 1$), with r no less than 0.09m, the estimate became worse off by a significant margin if the induced terms were ignored.

The vertical dashed lines in Figure 5.5 indicate the radii r below which the induced terms are regarded as negligibly small by our QP algorithm: they are $r = 0.01$m for $\epsilon = 0.1$, $r = 0.02$m for $\epsilon = 0.3$ and $r = 0.09$m for $\epsilon = 1$ respectively. Trying to fit these small induced terms below these threshold levels would cause overfitting and thus might be detrimental to performance. In view of the values which are chosen to reflect a fairly typical range of conditions encountered in both biological vision systems and artificial systems, our results suggest that the spherical camera system with a single viewpoint might not be adequate for modeling the large class of non-frontal eyes. This holds true especially if the separation between eyes are of medium length (e.g. for the vision systems of many large vertebrates and most non-miniaturized artificial systems), or the rotation is significant.

5.1.3 Effect of Calibration Errors for Lateral Eye

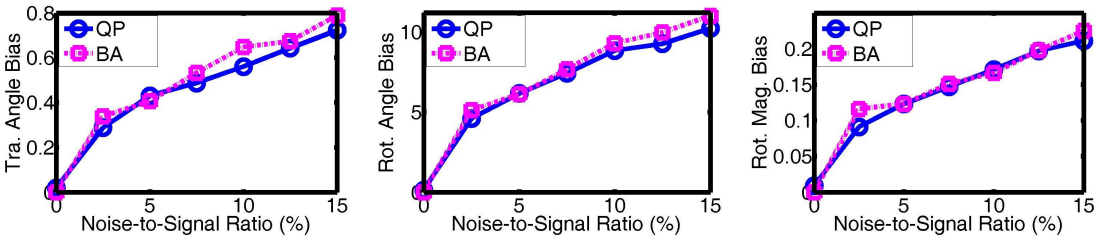
In real vision systems, there often exist imperfections in the construction of the compound eye. One kind of error is the imperfection in the spherical substrate of the com-

pound. This is especially pertinent in a biological compound eye system, where the head is not a perfect sphere. In other words, the radius r is different along different directions. The other kind of error stems from imperfections in the camera postures, where it could be difficult to perfectly align the optical axes of two opposite cameras. In this section, we conduct two experiments to test how both recovery methods (QP and BA) perform under such two kinds of errors. We used the same compound eye set-up in the previous section 5.1.1.3 and the same global motions as in Table 5.1.

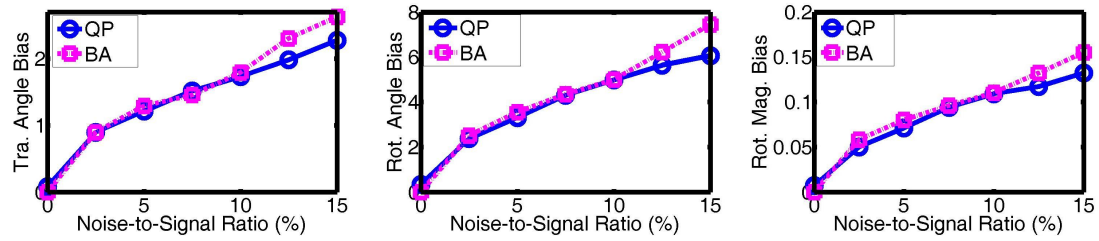
In the first experiment, we study the performance of the two estimation methods when the radius r is not constant, with variation of up to 50% being simulated across different pairs of ommatidia. In particular, r is set to be $r_0 + \Delta r_i$ where r_0 is the average value of the varying r and Δr_i is the random error added to the $i^{th}(i = 1 \dots 9)$ pair's separation with up to $0.5r_0$ variation. As before, the value of r_0 is assumed known for the BA.

On the whole, both methods are affected by the error in r (i.e. Δr_i) but to different extents. Compared to the identical performances seen in Figure 5.4, our method now gains a distinct advantage over the BA as shown in Figure 5.6. This could be attributed to the fact that our method is largely independent of the calibration parameter r; r only results in second and higher order residual terms which are negligible. Thus error in r has less an impact on our method's performance. In contrast, the BA's estimation algorithm depends more significantly on the fact that r is a known constant. Of course one can explicitly estimate these variations in r, but doing so would result in an algorithm that is much more complex and whose numerical performance is open to question.

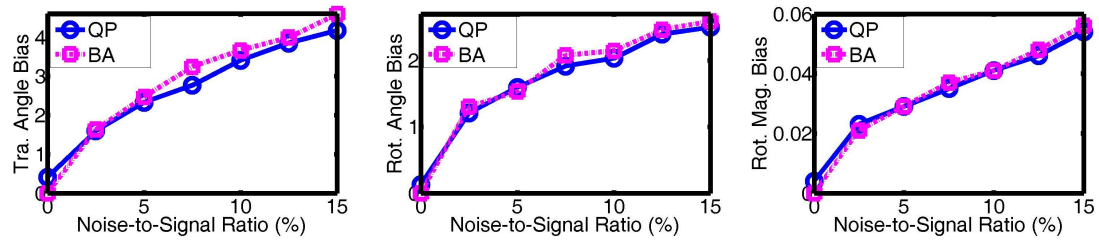
In the second experiment, we study how imperfections in the camera postures will affect the performance of these two methods. More specifically, consider the case where



(a) Estimation error when translation is dominant with ratio ε=5



(b) Estimation error when trans. and rot. are balanced with ratio ε=1

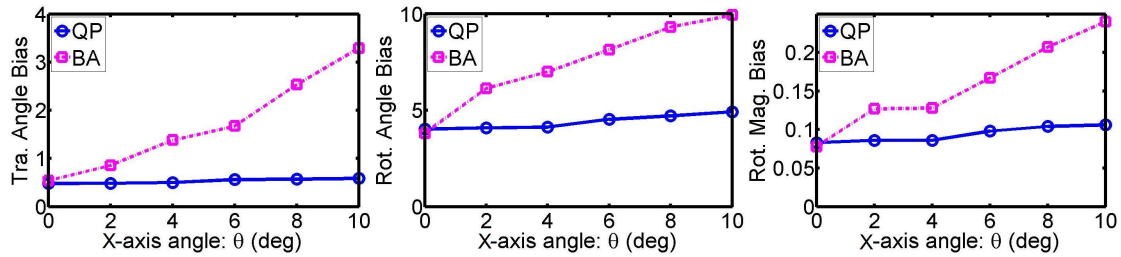


(c) Estimation error when rotation is dominant with ratio ε=0.2

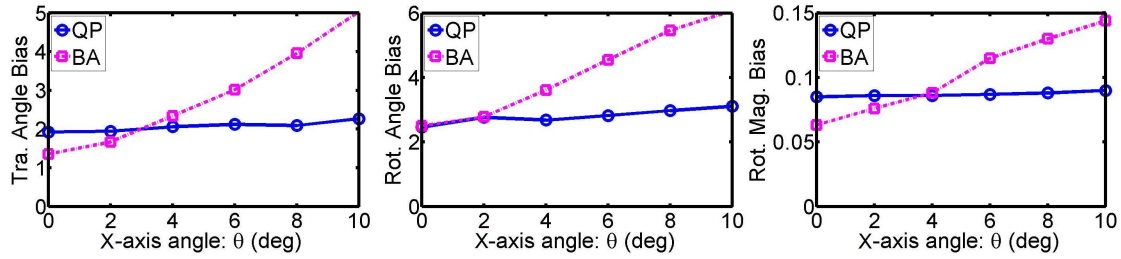
Figure 5.6: Motion recovery of the two methods in a compound eye placed on a non-perfect spherical substrate.

the two opposing cameras Cam1 and Cam2 of $50\circ$ FOV each in Figure 3.4 do not have their optical axes properly aligned. Without loss of generality, we can assume that the Cam2's viewing direction Z2 is aligned with the global Z-axis. Then the misalignment error of Cam1 can be modeled by the three consecutive rotations around the coordinate axes required to align the viewing direction of Cam1 with that of Cam2. Here, for simplicity, we only study the effect of misalignment error around the X-axis, which we model by the rotation angle θ. The flow input of the system was subject to a 10% isotropic noise.

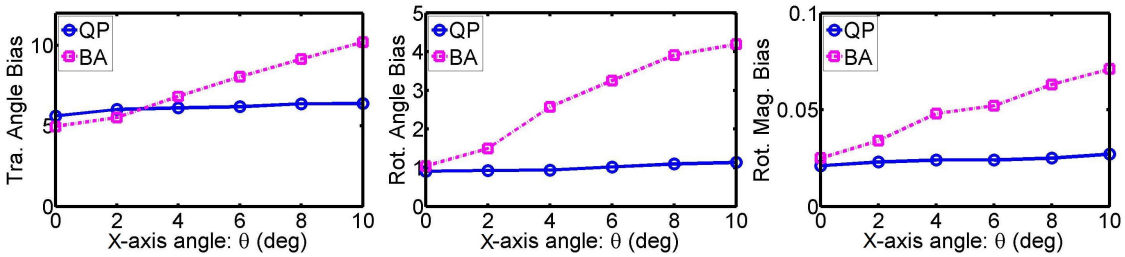
As the results in Figure 5.7 showed, under all conditions tested, our method outperformed the bundle adjustment considerably. It seemed that our method can tolerate misalignment error to a much greater extent than the bundle adjustment method. For



(a) Estimation error versus misalignment error θ, under ratio ε=5



(b) Estimation error versus misalignment error θ, under ratio ε=1



(c) Estimation error versus misalignment error θ, under ratio ε=0.2

Figure 5.7: Motion recovery of the two methods in a compound eye with misalignment error θ around the X-axis.

example, when misalignment error becomes substantial ($\theta = 10°$), the estimation error of our translation estimate was approximately $3°$ smaller under all ϵ. This superior performance of our algorithm can be attributed to the data normalization based on Total Least Squares, which is known to be suited for dealing with errors-in-variables model (misalignment error in this case).

5.1.4 Experiment on Real Image for Lateral Eye

In this experiment on real image, we mounted a lateral pair of cameras with $r = 0.12$m on a mobile robot. We used two Dragonfly cameras from Point Grey Research with $50°$ FOV each. The frame rate is 15 frames per second and the image size is 640×480

pixels. The lateral pair views an indoor scene with depths ranging from 2m to 5m. The robot moves on the floor with two degrees of freedom in the translation and one degree of freedom in the rotation. A picture of the scene taken by one of the cameras is shown in Figure 5.8. We tested three sets of global motions (see Figure 5.9 for details), with ε roughly equal to 0.2, 1 and 5 respectively. The robot moved approximately 2 meters. An average of 140 matching points were selected for the linear method while 850 feature points were selected for the BA. Figure 5.9 plots the bias in degrees for the two direction estimates and in $\%$ for the magnitude estimate. It shows that under all conditions, the estimation accuracy of the linear method was comparable to that of the BA.

Figure 5.8: An indoor scene for the real-image experiment.

5.2 Parallel Eye

5.2.1 Experiment on Range Image for Parallel Eye

Each camera has a 40° FOV and $h = 0.2$ m. 145 matching points were found for the linear method and 952 feature points were for BA.

Under all conditions tested in Fig.(5.10), the linear and BA methods produced almost identical results. With $\varepsilon = 1$, the linear method's performance is slightly inferior to the BA method. The maximal angular difference between the two was 1.6°, in the case of translation recovery. Decreasing ε from 1 to 0.2 expectedly improved rotation

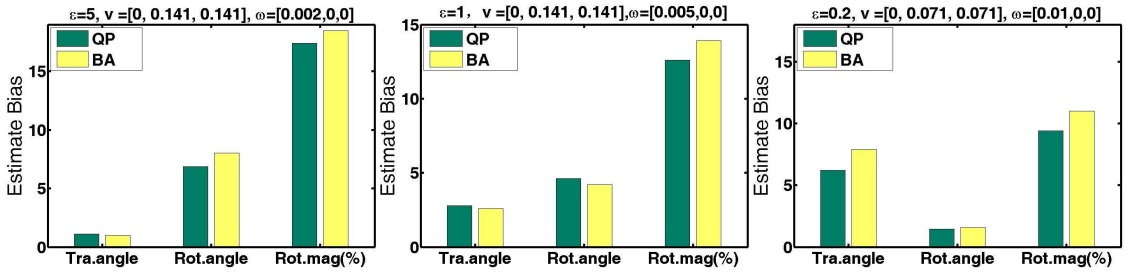


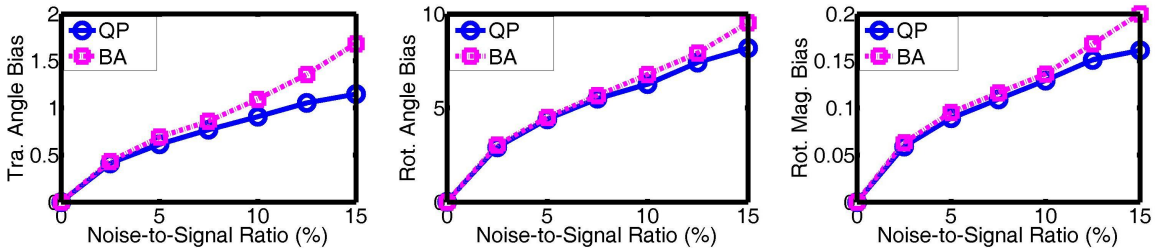
Figure 5.9: Motion recovery of the two methods using real image. The unit of translation **v** is m/s and the unit of rotation ω is rad/s. Bias plotted in degrees for the two direction estimates and in $\%$ for the magnitude estimate.

recovery while impairing translation recovery, as was observed in both methods. However compared to BA, our method produced a substantially improved rotation estimate with the translation estimate slightly worsened.

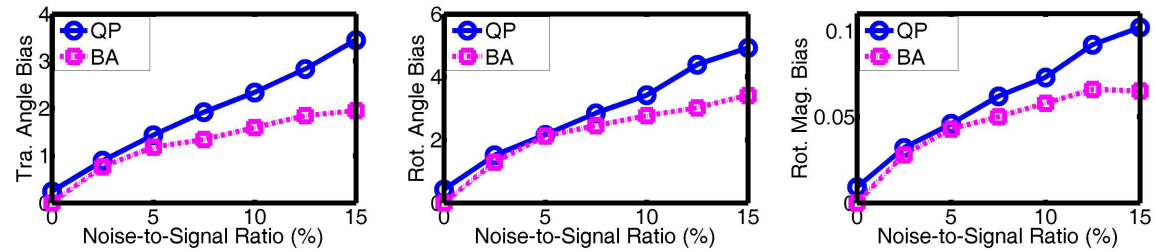
Raising ε from 1 to 5 generates a dominant translational flow. Both methods recovered translation accurately whereas the rotation estimates became worse off. However, BA's rotation recovery was affected more adversely: with 15% noise, our rotation estimate was worsened by $3.2°$, in contrast to a $6.5°$ increase in BA. Our better performance can be attributed to the post-translation step, where rotation estimation benefited substantially from a good translation estimate. Thus the potentially unfavorable condition for rotation recovery caused by the dominant translation was ameliorated by our method.

Overall, there is no clear winner among the two methods. While our linear method achieved the accuracy of the nonlinear BA, it required much less computational cost (about $\frac{1}{15}$ the computation time of BA), which is crucial for an agent with little computational resources.

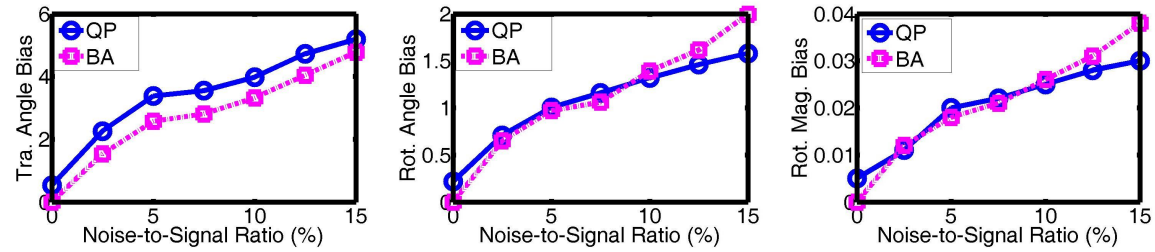
In real biological vision systems, there often exist a small amount of convergence between two eyes, i.e. stereos are no longer strictly parallel. Figure 5.11 shows a convergence angle formed by stereo cameras. While looking at near objects with large convergence angle, animals are more likely to use stereo cues. But we are interested in



(a) Estimate error when translation is dominant with ratio ε=5



(b) Estimate error when trans. and rot. are balanced with ratio ε=1



(c) Estimate error when rotation is dominant with ratio ε=0.2

Figure 5.10: Motion recovery of two methods. QP stands for quasi-parallax method, while BA denotes bundle adjustment. Starting from the left, the three columns depict the estimate errors in translation direction, rotation direction, and rotation magnitude respectively. Note that y-axis for different diagrams may have different scales.

the case of far vision, for human it means objects are beyond 1m with angle $< 5°$.

We conducted an experiment to show how the two recovery methods (QP and BA) perform when the camera pair are no longer parallel but has an unknown convergence angle instead. The convergence angle is modeled by a rotation angle θ around Y-axis. The flow input of the system was subject to a 10% isotropic noise. And the camera pair undergoes a global motion of balanced translation and rotation with ratio $\varepsilon = 1$.

As the results above showed, when the convergence angle increased, our method started to outperform the bundle adjustment. It seemed that our method can tolerate the convergence angle to a much greater extent than the bundle adjustment method.

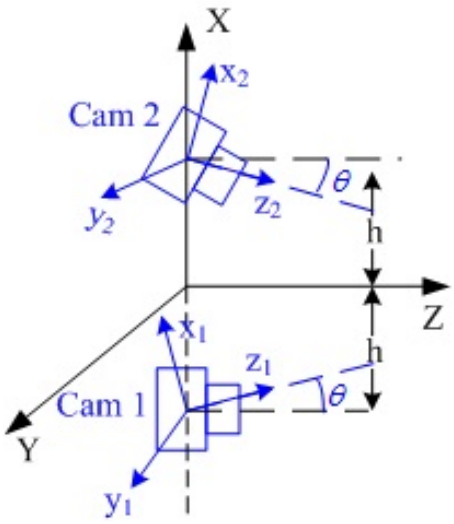


Figure 5.11: Stereo camera array with small covergence angle

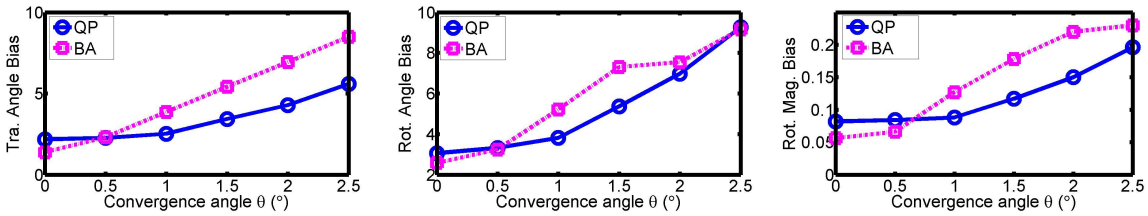


Figure 5.12: Motion recovery of both methods in a stereo eye of convergence angle θ around Y-axis

For example, when convergence angle becomes substantial ($\theta = 2.5°$ produces a $5°$ convergence angle), the estimation error of our translation estimate was approximately $4°$ smaller. Our robust performance results from the data normalization based on Total Least Squares, which can deal with errors-in-variables model).

In addition to the convergence angle, another imperfection in real biological vision system is the baseline error, i.e. the baseline can not be precisely measured and the baseline estimate has errors. We also want to study the effect of baseline estimate error on the performance of the two estimation methods. In the following experiment,the baseline h is set to be $h_0 + \Delta h$ where where h_0 is the baseline ground truth and Δh is the estimate error with up to $\pm 0.5 h_0$ variation. The camera array has a 10% isotropic noise and experiences a global motion with ratio $\varepsilon = 1$.On the whole, the estimate error

in h does not have a major effect on both methods. This could be due to the fact that the calibration parameter h only has results in second and higher order residual terms which are negligible.

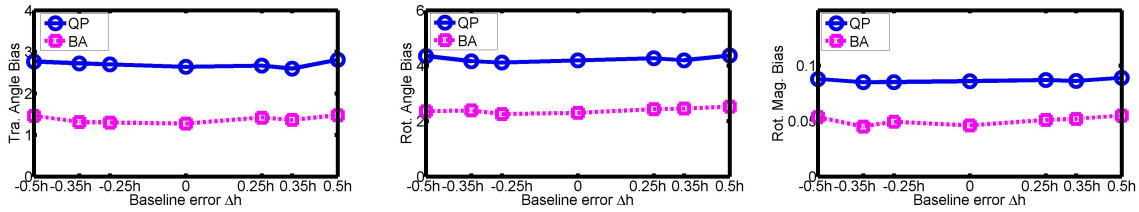


Figure 5.13: Motion recovery of both methods in a stereo eye with errors in baseline estimate

5.2.2 Experiment on Real Image for Parallel Eye

We mounted a parallel camera array with $h = 0.15$ m on a mobile robot. Each camera has a $50°$ FOV and views an indoor scene which is similar to Fig 5.8. We tested three sets of global motions, with ε roughly equal to 0.2, 1 and 5 respectively. Around 180 matching points were selected for the linear method and 750 feature points were selected for BA. Fig.(5.14) shows that under all conditions, the estimation accuracy of the linear method was comparable to that of BA.

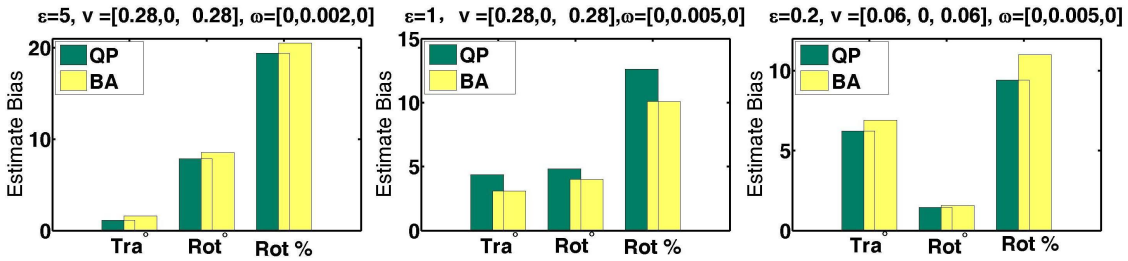


Figure 5.14: Motion recovery of both methods on real image. Bias plotted in degree for the two direction estimates and in % for the magnitude estimate.

In sum, under the three motion types we have tested, our linear method is comparable to Bundle adjustment in terms of recovery accuracy, while we are much faster. Interestingly, when the stereo pair is not perfectly parallel with a convergence angle,

the effect of such imperfection is lesser on the linear method. For the estimate error of baseline, both methods are immune from it.

Chapter 6

Experiments on Depth Reconstruction

As pointed out by the reliability analysis of depth recovery conducted in (Cheong and Xiang, 2001), when translation is coupled with rotation, with known or unknown intrinsic parameters, lateral motion is better than forward motion in terms of yielding ordinal depth information and other aspects of depth recovery. More specifically, for lateral movement, although Euclidean reconstruction is difficult, the resulting distortion in the structure possesses many nice properties. For example, in the case of calibrated motion, the distortion preserves ordinal depth. In contrast, for forward movement, whether calibrated or uncalibrated, depth information (even partial one) is hard to recover, except for those points close to the observer.

Specific motion-scene configurations can lead to the degeneracy problem (Hartley, 1997). For instance, this degenerate problem arises in a planar scene where all world points lie on a plane. For the discrete stereo set-up, this means there is no unique solutions for the fundamental matrix and consequently the projective reconstruction cannot be estimated. Moreover, configurations close to the degenerate ones are likely to lead to a numerically ill-conditioned estimation. In this chapter, we attempt to recover depths

for the case of a lateral eye system.

6.1 Depth Reconstruction Method

After recovering the ego-motion (U, V, W) of the lateral eye system, we proceed to reconstruct the depth scenes viewed by each camera. We assume the radius r to be known. Then, based on Equation 3.1, we can compute the motions of the two cameras in their own camera coordinate systems, from which we can recover depth from the following linear equation:

$$Z = \frac{(Wx - fU, Wy - fV) \cdot (n_x, n_y)}{(u - u_{rot}, v - v_{rot}) \cdot (n_x, n_y)} \tag{6.1}$$

Where (n_x, n_y) is an unit vector which specifies a direction. Here we choose to recover depth along the estimated epipolar direction, based on the intuition that the epipolar direction contains the strongest translational flow. It means that we first project optical flow along the direction emanating from the estimated FOE and then recover depth along that direction, i.e. $(n_x, n_y) = \frac{(Wx-fU, Wy-fV)}{\sqrt{(Wx-fU)^2+(Wy-fV)^2}}$, or in the case of $W = 0$ where the estimated FOE is at infinity, $(n_x, n_y) = -\frac{(U,V)}{\sqrt{U^2+V^2}}$

Denote the depth estimate at the i^{th} feature point as $\widehat{Z}_i$. To measure the performance of depth recovery method, we define the depth reconstruction error as $\frac{1}{N}\sum|\hat{Z}_i - Z_i)|$ where N is the total number of feature points.

6.2 Experiment of Lateral Set-up

We first conduct experiments on range image. We use a pair of lateral cameras with 50° of FOV each, and the radius r is 0.1m. In addition to the forest scene that was used in

Chapter 5, we also use the following office scene Fig 6.1 with depth ranging from 1m to 55m. After projecting the the 3D scene points onto the image planes, the 2D image points now have depths ranging from 2m to 7m. For the forest scene, the depth range is from 8m to 20m.

Figure 6.1: An office scene used. Intensity represents depth with distant object looking brighter. Regions with no range data appears black. The image plane of camera 1 is denoted by a white rectangle.

The depth reconstruction performance of both methods (linear quasi-parallax and bundle adjustment) were tested in various motion-scene configurations, with 10% of isotropic noise added. More specifically, the system executed three types of motions as shown in Table 6.1: lateral motion ($W = 0$), forward motion ($U = V = 0$) and general motion. The translation-to-rotation ratio ε was fixed at 1.

Table 6.1: Three types of global motions (lateral, forward and general) are executed to test the performance of depth reconstruction.

Motion Type	Translation: cm/s	Rotation: rad/s
Lateral	[2, 3, 0]	[0.004, 0.003, 0.002]
Forward	[0, 0, 7]	[0.004, 0.003, 0.002]
General	[1, 3, 2]	[0.004, 0.003, 0.002]

6.3 Experimental Results on Range Image

6.3.1 Perfect r

In this section, we evaluate the performance of depth reconstruction when the measurement of the radius r is prefect. Table 6.2 listed the numerical error of ego-motion esti-

	Tra. Dir	Tra. Mag	Rot. Dir	Rot. Mag	Forest	Office
Lateral QP	2.09°	18%	3.20°	10.0%	2.29 m	0.51 m
Lateral BA	1.73°	24%	3.65°	8.9%	1.82 m	0.63 m
General QP	1.93°	14%	2.96°	8.5 %	2.43 m	0.53 m
General BA	1.49°	12%	2.90°	7.2 %	2.57 m	0.48 m
Forward QP	1.24°	11%	1.56°	4%	3.27 m	1.05 m
Forward BA	1.26°	9%	1.37°	3%	3.7 m	1.21 m

Table 6.2: Recovery results of our linear quasi-parallax method (QP) and Bundle adjustment (BA). Starting from the left, the columns represent errors of translation direction in degree, errors of translation magnitude, errors of rotation direction, errors of rotation magnitude, reconstruction error of forest and reconstruction error of office.

mation, along with the depth reconstruction error. These results were further compared against those of Bundle adjustment. For brevity, we will only show the reconstruction results of one camera and here we choose camera 1. In Figure 6.2, the depth maps of the two scenes were depicted using a color coding scheme: the colorbars showed that cool colors such as deep blue represented closer feature points, while warm colors such as red means that the points were far away. Note that the mappings between the color and the depth were different for the two scenes, so that the plots were easier to read.

For both scenes, we see that under all the motion types, both methods yielded similar results and the depth reconstructions were quite good. More specifically, depth orders were preserved for most of the feature points, though a precise metric Euclidean reconstruction was still difficult, as can be seen from the last two columns of Table 6.2. Among the three different types of global motion, lateral motion generated a depth map with the highest accuracy, despite the fact that its ego-motion estimate is the least ac-

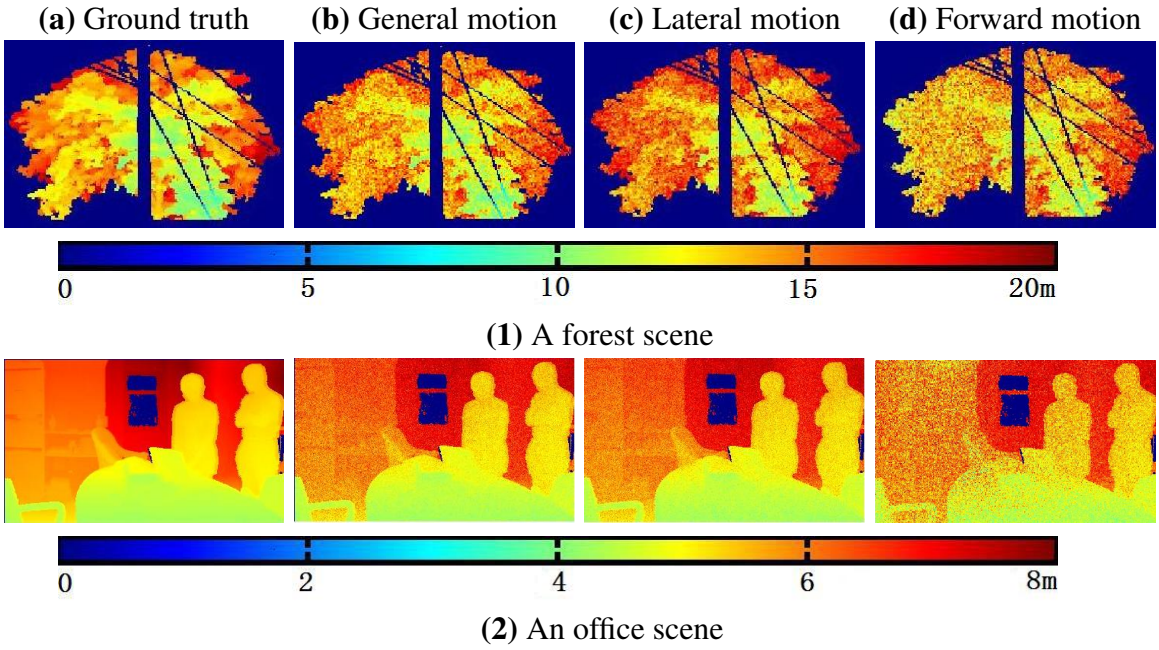


Figure 6.2: Recovered depthmaps for various motion-scene settings. The first row **(1)** shows a forest scene and the second row **(2)** shows an indoor office. The columns represent respectively **(a)** ground truth depthmap, **(b)** recovered depthmap under general motion, **(c)** depthmap under lateral motion, **(d)** depthmap under forward motion. Note that each scene has its own individual color mapping scheme.

curate. This confirms to the previous findings that lateral motion is conducive to depth recovery, although in such case the bas-relief ambiguity is difficult to resolve and causes errors in the motion estimates. In contrast, forward motion is not well suited for depth recovery but good for ego-motion recovery.

6.3.2 Erroneous r

In real applications, it is not easy to get an accurate measurement of the radius r. We model this case by introducing a random error Δr into the true radius r_0: the radius r is $r_0 + \Delta r$ and Δr has up to $50\% r$ variation. In this experiment, we test the reconstruction of the forest scene, knowing only r_0.

Table 6.3 clearly shows that our algorithm is quite immune from error in r. According to Equation 6.1, there are two ways that Δr affects the depth recovery. First, it is

through the global translation parameters. As mentioned in Sect 5.1.3, U, V, W can be recovered accurately even with Δr present. The second way is through the terms of $\alpha\Delta r$ and $\beta\Delta r$. Normally these terms are much smaller than U and V in magnitude. Thus on the whole, the effect of Δr is negligible.

Table 6.3: Three types of global motions (lateral, forward and general) are executed to test the performance of depth reconstruction.

	Δr=0	$\Delta r = 0.2r_0$	$\Delta r = 0.5r_0$
Lateral	2.29 m	2.37 m	2.51 m
General	2.43 m	2.42 m	2.45 m
Forward	3.27 m	3.27 m	3.29 m

6.4 Experimental Results on Real Image

As for the experiment on real-image, the ground truth of the depth map is not available. Instead, we checked the number of depth orders that were recovered correctly. Though it is not a comprehensive test, it serves as a good performance measure of the ordinal depth recovery. We manually picked and divided feature points into foreground points and background points. In all, we picked 530 foreground points and 662 background points (some are shown in Figure 6.3); and then compared the recovered depth of each foreground point with that of each background point. The former should be smaller than the latter if the depth order is preserved.

As mentioned above, we used the number of correct depth orders to measure the performance of depth recovery. As the results in Table 6.4 shows, the rate of correct depth order was quite high in both methods, even in the unfavorable condition of forward motion.

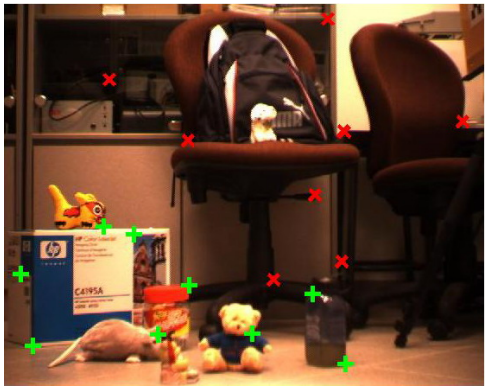


Figure 6.3: The indoor scene from the real-image experiment, with foreground points marked as green + and background points as red x.

Table 6.4: Depth reconstruction of two methods in the real-world experiment. The table show the percentage of points whose depth orders were preserved.

Correct %	Forward $V = [0, 0, 0.28]^T, \omega = [0.005, 0, 0]^T$	General $V = [0, 0.141, 0.141]^T, \omega = [0.005, 0, 0]^T$
QP	52.3%	77.6%
BA	56.1%	79.8%

Chapter 7

Conclusion

Having eyes that look at diametrically opposite parts of the world is a common form of visual field layout found throughout the biological world. Such form of eye arrangement is realized in many animals including the lateral eyes of vertebrates such as birds and fishes, and invertebrates such as insects with their compound eyes. Though theories have been advanced about how these animals exploit this special eye topography to accomplish complex visuomotor tasks with limited brain power, these theories lack computational underpinning.

Part of this thesis addresses this gap by investigating computationally how the ego-motion recovery problem can gain maximal benefit from the opportunities afforded by this particular form of eye arrangement. We showed that the ego-motion can be solved by making full use of the special eye structure without resorting to complex and computationally expensive algorithms. We proposed a linear method based on the notion of quasi-parallax. It makes use of a matching pair of diametrically opposite visual rays to directly recover the heading direction, without any need of finding correspondences nor requiring nonlinear optimization.

Our method recovers the translation and rotation separately by looking at different ensembles of projection rays. The quasi-parallax motion field contains terms primarily arising from global translation, save for the residual terms caused by induced translation. Therefore, information pick-up for the translation is enhanced. The accuracy of the translation estimate is further improved by a small iterative step that computes the induced terms. Given this translation estimate, the rotation is recovered from a pencil of visual rays using the individual epipolar constraints of each camera. As a consequence of this two-stage process that selects different and appropriate aspects of the visual rays, both the translation and rotation can be recovered well even under adverse conditions, such as dominant translational or rotational flow coupled with a high level of noise.

Statistically, the Fisher information matrix corroborates our conjecture that the quasi-parallax method is more effective in resolving the bas-relief ambiguity than the BA algorithm, especially under small and moderate field of views. This is also verified by the experimental results obtained under a pair of lateral eyes with narrow FOV. For other scenarios such as wide FOV, cameras arranged in resemblance of a compound eye, real images with non-ideal feature distribution, our method achieved a comparable performance compared to that of the BA algorithm. We also showed that our method is robust against imperfection in the construction of the spherical substrate of the compound eye. Variation up to 50% in the radius r and misalignment error up to 10° resulted in graceful deterioration of the performance for our method, whereas the BA method showed a greater drop in accuracy. This is possibly because the BA method relied more significantly upon the fact that r is a constant; besides our data normalization scheme based on Total Least Squares is well-suited to dealing with such errors-in-variables model. It is also worth emphasizing that our method requires much less computational cost and

calibration efforts, a significant advantage in any visual system with a need for rapid visuomotor coordination.

We also present schemes that exploits eyes of the same viewing directions, namely the Aranead eye and parallel pair of eyes. Such eyes are inspired by predatory animals which tend to have the eyes facing in the same direction. A common view is that this arrangement is to leverage the stereoscopic depth perception. Yet we explore the other possibility that for animals with restricted neural power, the general function of binocularity is concerned with optic flow-fields. The schemes obtain motion parallax from subtracting optical flows at the matching points. The resulting motion parallax is proven to facilitate the translation recovery and subsequently the motion recovery.

Once the motion parameters were recovered, we proceeded to linearly reconstruct the depth maps for both indoor and outdoor scenes. It was shown that given accurate motion estimates, the depth reconstruction was fast and reliable, capable of being used in real-time tasks.

References

[Adiv1989] Adiv, Gilad. 1989. Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field. *PAMI*, 11(5):477–489.

[Anandan and Irani2002] Anandan, P. and M. Irani. 2002. Factorization with uncertainty. *IJCV*, 49(2-3):101–116.

[Argyros and Orphanoudakis1997] Argyros, A.A. and S.C. Orphanoudakis. 1997. Independent 3d motion detection based on depth elimination in normal flow fields. *Proc. IEEE CVPR*, pages 672–677.

[Ayache and Faugeras1989] Ayache, N. and O.D. Faugeras. 1989. Maintaining representations of the environment of a mobile robot. *IEEE Trans. Robotics and Automation*, 5(5):804–819.

[Baker, Ogale, and Fermüller2004] Baker, P., A.S. Ogale, and C. Fermüller. 2004. The Argus eye, a new tool for robotics. *IEEE Robotics and Automation Magazine*, 11(4):31–38.

[Blanke, Nalbach, and Varju1997] Blanke, H., H.-O. Nalbach, and D. Varju. 1997. Whole-field integration, not detailed analysis, is used by the crab optokinetic system to separate rotation and translation in optic flow. *Journal of Comparative Physiology A*, 181:383–392.

[Bradshaw and Rogers1993] Bradshaw, M. and B. J. Rogers. 1993. Subthreshold interactions between binocular disparity and motion parallax. *Investigative Ophthalmology and Visual Science*, 33 (Suppl. 4):1332.

[Chahl and Srinivasan1997] Chahl, J.S. and M.V. Srinivasan. 1997. Reflective surfaces for panoramic imaging. *Applied Optics*, 36(31):8275–8285.

[Cheong and Xiang2001] Cheong, Loong Fah and Tao Xiang. 2001. Characterizing depth distortion under different generic motions. *International Journal of Computer Vision*, 44(3):199–217.

[Clark and Yuille1994] Clark, J. and A. Yuille. 1994. *Data fusion for sensory information processing.* Kluwer, Boston.

[Coombs and Roberts1993] Coombs, D. and K. Roberts. 1993. Centering behavior using peripheral vision. *CVPR*, pages 440–451.

[Davies and Green1994] Davies, M. N. O. and P. R. Green. 1994. Multiple sources of depth information: an ecological approach. In *Perception and motor control in birds: an ecological approach*, pages 339–356. Springer.

[Duparre and et al.] Duparre, J. and et al. Artificial compound eyes: different concepts and their application for ultraflat image acquisition sensors. In *Proc. SPIE.*

[Fermüller and Aloimonos2000] Fermüller, C. and Y. Aloimonos. 2000. Observability of 3d motion. *IJCV*, 37(1):43–63.

[Franz and et al1998a] Franz, M.O. and et al. 1998a. Learning view graphs for robot navigation. *Autonomous Robots*, 5:111–125.

[Franz and et al1998b] Franz, M.O. and et al. 1998b. Where did I take that snapshot? Scene-based homing by image matching. *Biological Cybernetics*, 79:191–202.

[Grosso, Sandini, and Tistarelli1989] Grosso, E., G. Sandini, and M. Tistarelli. 1989. 3-D Object Reconstruction Using Stereo and Motion. *IEEE Trans. Systems, Man and Cybernetics*, 19(6):1465–1476.

[Grosso and Tistarelli1995] Grosso, E. and M. Tistarelli. 1995. Active/Dynamic Stereo Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(11):11171128.

[Haag and Borst2001] Haag, J. and A. Borst. 2001. Recurrent network interactions underlying flow-field selectivity of visual interneurons. *J Neuroscience*, 21:5685–5692.

[Hartley1997] Hartley, R. 1997. In defense of the eight-point algorithm. *PAMI*, 19(6):580–593.

[Hartley and Zisserman2000] Hartley, R. and A. Zisserman. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[Heeger and Jepson1992] Heeger, D. J. and A. D. Jepson. 1992. Subspace methods for recovering rigid motion: I. algorithm and implementation. *IJCV*, 7(2):95–117.

[Ho and Chung2000] Ho, P. K. and R. Chung. 2000. Stereo-Motion with Stereo and Motion in Complement. *PAMI*, 22(2):215–220.

[Hornsey and et al.2004] Hornsey, R. and et al. 2004. Electronic compound eye image sensor: construction and calibration. In *Proc. SPIE*, volume 5301, pages 13–24.

[Huber and Bülthoff1998] Huber, S.A. and H.H. Bülthoff. 1998. Simulation and robot implementation of visual orientation behaviors of flies. In *From Animals to Animats 5, Proc. of SAB'98*, pages 77–85.

[Huguet and Devernay2007] Huguet, F. and F. Devernay. 2007. A Variational Method for Scene Flow Estimation from Stereo Sequences. *ICCV*.

[Ibbotson1991] Ibbotson, M.R. 1991. Wide-field motion sensitive neurons tuned to horizontal movement in the honeybee *Apis mellifera*. *Journal of Comparative Physiology A*, 168:91–102.

[Jeong, Kim, and Lee2006] Jeong, K., J. Kim, and L.P. Lee. 2006. Biologically inspired artificial compound eyes. *Science*, 321:557–561.

[Johnston, Cumming, and Lany1994] Johnston, E. B., B. G. Cumming, and M. S. Lany. 1994. Integration of stereopsis and motion shape cues. *Vision Research*, pages 2259–2275.

[Kern1998] Kern, R. 1998. Visual position stabilization in the hummingbird hawk moth, *Macroglossum stellatarum* L. II. Electrophysiological analysis of neurons sensitive to wide-field image motion. *Journal of Comparative Physiology A*, 182:239–249.

[Kim, Jeong, and Lee2005] Kim, J., K. Jeong, and L. P. Lee. 2005. Artificial ommatidia by self-aligned microlenses and waveguides. *Optical Letter*, 30:5–7.

[Kriegman, Triendl, and Binford1989] Kriegman, D.J., E. Triendl, and T.O. Binford. 1989. Stereo vision and navigation in buildings for mobile robots. *IEEE Trans. Robotics and Automation*, 5(6):792– 803.

[Land and Nilsson2006] Land, M. F. and D-E. Nilsson. 2006. General purpose and special purpose visual systems. In *Invertebrate vision*, pages 167–210. Cambridge University Press.

[Lee and Huang2000] Lee, A. B. and J. Huang. 2000. Brown range image database. http://www.dam.brown.edu/ptg/brid/index.html.

[Leedan and Meer2000] Leedan, Y. and P. Meer. 2000. Heteroscedastic regression in computer vision: Problems with bilinear constraint. *IJCV*, 37(2):127 – 150.

[Leonard, J.I.Simpson, and Graf1988] Leonard, C.S., J.I.Simpson, and W. Graf. 1988. Spatial organization of visual messages of the rabbit's cerebellar flocculus. *Journal of Neurophysiology*, 60:2073–2090.

[Li and Duncan1993] Li, L. and J.H. Duncan. 1993. 3-d translational motion and

structure from binocular image flows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(7):657–667.

[Li and Sclaroff2008] Li, R. and S. Sclaroff. 2008. Multi-scale 3d scene flow from binocular stereo sequences. *Computer Vision and Image Understanding*, 110(1):75–90.

[Lim and Barnes2007] Lim, J. and N. Barnes. 2007. Estimation of the epipole using optical flow at antipodal points. *OMNIVIS*, pages 1–6.

[Lim and Barnes2008] Lim, J. and N. Barnes. 2008. Directions of egomotion from antipodal points. *CVPR*.

[Longuet-Higgins and Prazdny1980] Longuet-Higgins, H. C. and K. Prazdny. 1980. The interpretation of a moving retinal image. *Proc. Royal Soc. London, B,*, 208:385–397.

[Ma, Kosecka, and Sastry2001] Ma, Yi, Jana Kosecka, and Shankar Sastry. 2001. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, 44(3):219–249.

[Martin and Katzir1999] Martin, G. R. and G. Katzir. 1999. Visual field in short-toed eagles *Circaetus gallicus* and the function of binocularity in birds. *Brain Behavior Evolution*, 53:55–66.

[McFadden1994] McFadden, S. A. 1994. Binocular depth perception. In *Perception and motor control in birds: an ecological approach*. Springer, pages 54–73.

[Nawrot and .1993] Nawrot, M. and R. Blake . 1993. On the perceptual identity of dynamic stereopsis and kinetic depth. *Vision Research*, 33:1561– 1571.

[Neumann and et al.2004] Neumann, J. and et al. 2004. Compound eye sensor for 3D

ego-motion estimation. *IEEE International Conference on Intelligent Robots and Automation*, pages 3712 – 3717.

[Ogata, Ishida, and Sasano1994] Ogata, S., J. Ishida, and T. Sasano. 1994. Optical sensor array in an artificial compound eye. *Optical Engineering*, 33:3649–3655.

[Okutomi and Kanade1993] Okutomi, M. and Takeo Kanade. 1993. volume 15.

[P. and M.A.1991] P., Alasubramanyan and Snyder M.A. 1991. The p-field: A computational model for binocular motion processing. *Proc. IEEE CVPR*, pages 115–120.

[Pless2004] Pless, R. 2004. Camera cluster in motion: Motion estimation for generalized camera designs. *IEEE Robotics and Automation Magazine*, 11(4):39–44.

[Pons, Keriven, and Faugeras2007] Pons, JP, R. Keriven, and O. Faugeras. 2007. Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score. *IJCV*, 72(2):179–193.

[Shi, Shu, and Pan1994] Shi, Y.Q., C.Q. Shu, and J.N. Pan. 1994. Unified optical flow field approach to motion analysis from a sequence of stereo images. *Pattern Recognition*, 27(12):1577–1590.

[Srinivasan and et al.1991] Srinivasan, M.V. and et al. 1991. Range perception through apparent image speed in freely-flying honeybees. *Visual Neuroscience*, 6:519–535.

[Srinivasan and et al.1996] Srinivasan, M.V. and et al. 1996. Honeybee navigation en route to the goal: Visual flight control and odometry. *Journal of Experimental Biology*, 199:237–244.

[Srinivasan, Zhang, and Chahl2001] Srinivasan, M.V., S. Zhang, and J.S. Chahl. 2001. Landing strategies in honeybees, and possible applications to autonomous airborne vehicles. *Biological Bulletin 200*, pages 216–221.

[Stein and Shashua1998] Stein, G. P. and A. Shashua. 1998. Direct estimation of motion and extended scene structure from a moving stereo rig. *CVPR*, pages 211–218.

[Stewenius and Astrom2004] Stewenius, H. and K. Astrom. 2004. Structure and motion problems for multiple rigidly moving cameras. *ECCV*, pages 252–263.

[Strecha and Gool2002] Strecha, C. and L. V. Gool. 2002. Motion-stereo integration for depth estimation. In *ECCV*.

[Sturm2005] Sturm, P. 2005. Multi-view geometry for general camera models. In *CVPR*, volume 1, pages 206–212.

[Symosek et al.1990] Symosek, P.F., B. Bhanu, S. Snyder, and B. Roberts. 1990. Motion and binocular stereo for passive ranging. In *IUW*, pages 358–363.

[Tao, Sawhney, and Kumar2001] Tao, H., H S. Sawhney, and R. Kumar. 2001. Dynamic depth recovery from multiple synchronized video streams. *CVPR*, pages 118–124.

[Thomas and Simoncelli1994] Thomas, I. and E. Simoncelli. 1994. Linear structure from motion. In *Technical report*. IRCS, University of Pennsylvania.

[Tittle and Braunstein1993] Tittle, J. S. and M. L. Braunstein. 1993. Recovery of 3-d shape from binocular stereopsis and structure from motion. *Perception and Psychophysics*, 54:157–169.

[Tomasi and Kanade1992] Tomasi, Carlo and Takeo Kanade. 1992. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154.

[Toscani and Faugeras1986] Toscani, G. and O. D. Faugeras. 1986. Structure and motion from two noisy perspective images. *Proceedings of IEEE Conference on Robotics and Automation*, pages 221–227.

[Triggs and et al.2000] Triggs, B. and et al. 2000. Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory and Practice*. Springer-Verlag, pages 298–375.

[Trucco and Verri1998] Trucco, E. and A. Verri. 1998. *Introductory Techniques for 3-D computer Vision*. Prentice Hall.

[Tsao and et al.1997] Tsao, A.T. and et al. 1997. Ego-motion estimation using optical flow fields observed from multiple cameras. *CVPR*.

[Vedula et al.2005] Vedula, S., S. Baker, P. Rander, R. Collins, and T. Kanade. 2005. Three-dimensional scene flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):475–480.

[Warrant2004] Warrant, E. 2004. Vision in the dimmest habitats on earth. *J Comp Physiol A*, (190):765–789.

[Waxman and Duncan1986] Waxman, A. M. and J. H. Duncan. 1986. Binocular image flow: steps toward stereo-motion fusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:715– 729.

[Weng, Ahuja, and Huang1993] Weng, Juyang, Narendra Ahuja, and Thomas S. Huang. 1993. Optimal motion and structure estimation. *PAMI*, 15(9):864–884.

[Williams, Isard, and MacCormick2005] Williams, O., M. Isard, and J. MacCormick. 2005. Estimating disparity and occlusions in stereo video sequences. In *CVPR*, volume 2, pages 250–257.

[Wylie and Frost1999] Wylie, D. R.W. and B.J. Frost. 1999. Responses of neurons in the nucleus of the basal optic root to translational and rotational flowfields. *Journal of Neurophysiology*, 81(2):267–276.

[Xiang and Cheong2003] Xiang, T. and L.F. Cheong. 2003. Understanding the behavior of SFM algorithms:a geometric approach. *IJCV*, 51(2):113–117.

[Zhang and Negahdaripour2008] Zhang, H and S Negahdaripour. 2008. Epiflow—a paradigm for tracking stereo correspondences. *Computer Vision and Image Understanding*, 111:307–328.

[Zhang1995] Zhang, Z.Y. 1995. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *PAMI*, 17(12):1222–1227.