Int J Comput Vis (2009) 84: 21–39
DOI 10.1007/s11263-009-0226-0

Linear Quasi-Parallax SfM Using Laterally-Placed Eyes

Chuanxin Hu · Loong Fah Cheong

Received: 24 January 2008 / Accepted: 23 February 2009 / Published online: 21 March 2009

Abstract 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 quasi-parallax 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.

Keywords Ego-motion estimation based on optical flow · Quasi-parallax terms · Lateral camera pairs · Compound eyes

C. Hu (✉)
Search Technology Center, Microsoft Research Asia, 4F Beijing Sigma Center, 49 Zhichun Road, Beijing, China
e-mail: chhu@microsoft.com

L.F. Cheong
Electrical and Computer Engineering, National University of Singapore, Block E4, Level 5, 4 Engineering Drive 3, Singapore, Singapore
e-mail: eleclf@nus.edu.sg

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 paper, 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.

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 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 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 et al. 2001) and view-based navigation (Franz 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 et al. 2004; Tsao 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 et al. 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.

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, 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 paper, 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 organization of this paper is as follows. Section 2 reviews the biological and the biomimetic robotics literature and discusses the wide variety of eye designs found in the vertebrates and the invertebrates. Section 3 seeks to situate our proposed method in the vast Structure from Motion literature, relating our work to other research efforts and paradigms. Section 4 discusses the approach and contribution of this paper and lists those systems to which our approach applies. Section 5 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. Section 6 studies the inherent ambiguity in our formulation using Fisher Information matrix and compares it against the "gold standard" solution obtained by the Bundle Adjustment method. Section 7 reports a set of experiments 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. Finally this paper ends with discussion and conclusion in Sect. 8.

2 Biological Vision Systems for Motion Estimation

2.1 Vertebrate Eyes: The Case of Birds

In birds, the second largest group of vertebrates after fishes, the assumption of stereopsis has been questioned. 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 instead have to do with motion processing.

Many species 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. In these species, maximum binocular field width is only about 10°, and the binocular field may be only 5° wide at the horizontal. Even 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. Such interspecific comparison of the available binocular field casts doubts on the utility of stereopsis for general scene perception and locomotion, except for visuomotor tasks involving close objects such as the bills.

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.

2.2 Invertebrate Eyes: The Case of Insects

Invertebrates have the greatest variety of eye types, and probably the 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. 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!

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 (FOV) 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 (e.g, crab: Blanke et al. 1997; bee: Ibbotson 1991; moth: Kern 1998; fly: Haag and Borst 2001). For instance, the well-studied "HS" cells of the fly pool optic flow from both eyes to estimate rotation and provide information for optomotor control and trajectory stabilization. Anyone who has ever observed blowflies chasing each other will be conversant with the breath-taking aerial acrobatics these tiny animals can produce, often among dense and complex environment.

3 Review of Computational Literature

Ego-motion estimation methods typically consider a monocular camera in motion or equivalently, camera views ob-

tained from multiple positions. These include the classical two-view methods (linear subspace Heeger and Jepson 1992, eight-point algorithm Hartley 1997) and the multiple view approaches (multiple view tensor Hartley and Zisserman 2000). The two-view methods suffer from the bas-relief ambiguity between translation and rotation, which cannot be removed by any statistical schemes, according to Xiang and Cheong (2003). Multiple-frame algorithms attempt to overcome this inherent limitation by incorporating redundancy from multiple frames but other difficulties set in such as tracking features over multiple frames.

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 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 measured flows and estimated flows, BA yields maximum likelihood estimates given a Gaussian image noise.

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.

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 (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 (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 et al. 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 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. (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 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, 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, 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, 2008; Thomas and Simoncelli 1994) and our work will be highlighted in later sections.

4 Linear SfM Based on Quasi-Parallax

This paper 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 Sect. 4.2). 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.

4.1 Outline of Our Approach

The main contribution of our paper 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 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 this paper 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.

4.2 Physical Realization

The following paragraphs give examples of visual systems to which our method can be applied.

4.2.1 Artificial Compound Eye

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.

Such technology allows ommatidia to be arranged normal to a sphere, more faithful to their natural counterparts. Hornsey 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 et al. (2005). This was followed by a 3D compound eye with self-aligned waveguides and individual microlens units on a spherical surface by Jeong et al. (2006) as shown in Fig. 1. The ommatidia were arranged along a hemispherical polymer dome such that each points to a different direction, allowing for a wide FOV, similar to that of the natural eye. Neumann 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. All these visual systems, by virtue of their wide FOVs, possess matching visual rays.

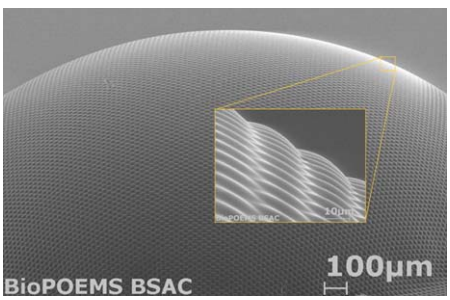


Fig. 1 An artificial compound eye fabricated by the biologically inspired 3D optical synthesis method from Jeong et al. (2006)

4.2.2 Lateral Eyes

The visual system of many vertebrates has eyes that are not frontally placed. Each eye views different parts of the world with little or no overlap in the visual fields. 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.

4.2.3 Conical Mirror

The conical mirror camera has become quite popular in biomimetic vision systems since it provides a 2D model of the almost omnidirectional insect eye, and it is relatively easy to construct (Chahl and Srinivasan 1997; Franz et al. 1998b; Huber and Bülthoff 1998). Referring to Fig. 2, if the upper limiting ray l_u is above the horizon (this is the case if $\alpha > 90°$ and $R > -h\cos\alpha\tan\frac{\alpha}{2}$), 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.

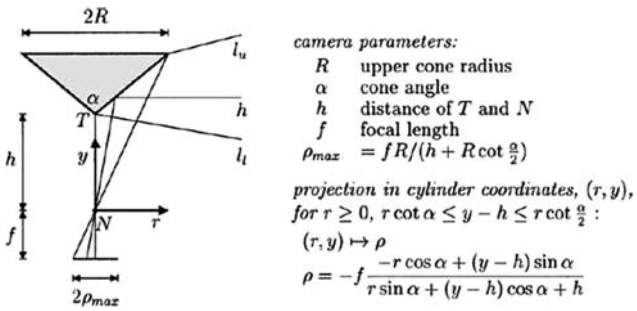


Fig. 2 Imaging geometry for a conical mirror camera from Chahl and Srinivasan (1997). N, camera nodal point; T, tip of conical mirror; l_u, upper limiting ray; h, horizontal ray; l_1, lower limiting ray. The conical mirror camera allows for capturing omnidirectional images without rotating a camera. A ring-shaped visual field between l_1 and l_u is mapped to a circular disk with radius $\rho_{\max}$ in the image plane. The visual field contains the horizon if $\alpha > 90°$ and $R > -h\cos\alpha\tan\frac{\alpha}{2}$

5 Technical Details of our Approach

5.1 Prerequisites

Figure 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 Sect. 5.3.

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 $\{O\text{-}XYZ\}$. The cth camera's translation $\mathbf{v}_c$ and rotation ω_c are related to the global motion by the following respectively:

$$\mathbf{v}_c = \mathbf{R}_c^T \cdot (\omega \times \mathbf{T}_c + \mathbf{v}), \quad \omega_c = \mathbf{R}_c^T \cdot \omega \tag{1}$$

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

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

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

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

$$\begin{aligned} \mathbf{v}_1 &= (U + \beta r, V - \alpha r, W)^T, \quad \omega_1 = (\alpha, \beta, \gamma)^T \\ \mathbf{v}_2 &= (-U + \beta r, V + \alpha r, -W)^T, \quad \omega_2 = (-\alpha, \beta, -\gamma)^T \end{aligned} \tag{2}$$

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:

$$\begin{aligned} u &= \frac{u^{tr}}{Z} + u^{rot} \\ &= \frac{Wx - fU}{Z} + \frac{\alpha xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \\ v &= \frac{v^{tr}}{Z} + v^{rot} \\ &= \frac{Wy - fV}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x \end{aligned} \tag{3}$$

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:

$$uv^{tr} - vu^{tr} = u^{rot}v^{tr} - v^{rot}u^{tr} \tag{4}$$

Fully expanding (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.

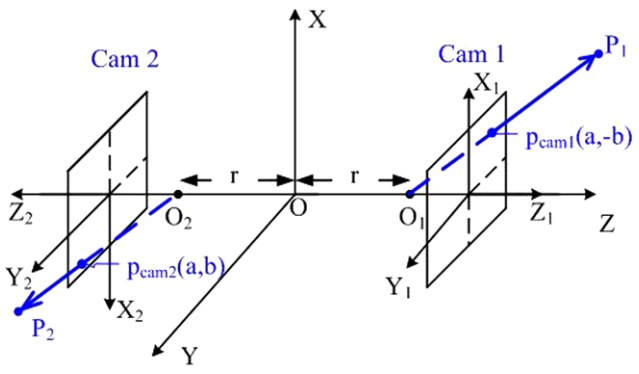


Fig. 3 Configuration of the laterally-placed pair of cameras. $\mathrm{p}_{\mathrm{cam1}}$ and $\mathrm{p}_{\mathrm{cam2}}$ form a matching pair of points

5.2 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 Fig. 3, $\mathbf{p}_{\mathrm{cam1}}$ and $\mathbf{p}_{\mathrm{cam2}}$ are a pair of matching points. $\mathbf{p}_{\mathrm{cam1}}$ is projected from visual ray O_1P_1 and $\mathbf{p}_{\mathrm{cam2}}$ is projected from O_2P_2. If the image coordinate of $\mathbf{p}_{\mathrm{cam1}}$ is $(a, -b)$, it is evident that $\mathbf{p}_{\mathrm{cam2}}$ should lie at the position of (a, b). In general, we use $\mathbf{p}_{\mathrm{cam1}} = (x^i, y^i)$ and $\mathbf{p}_{\mathrm{cam2}} = (x^i, -y^i)$ to represent the ith matching pair. In the cth $(c = 1, 2)$ camera, denote (u_c^i, v_c^i) as its optical flow measured at the ith 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.

5.2.1 Stage 1: Recovering the Global Translation

At the matching points, we substitute the respective camera rotations in (2) into (3). It is clear that the rotational flows at the pair of matching points are identical in magnitude:

$$u_2^{rot} = u_1^{rot} \triangleq u^{rot}, \qquad v_2^{rot} = -v_1^{rot} \triangleq -v^{rot} \tag{5}$$

With these two equalities in hand, we subtract the respective epipolar constraints (4) of camera 1 and camera 2 and obtain:

$$\begin{aligned} &u_1 v_1^{tr} - u_2 v_2^{tr} - v_1 u_1^{tr} + v_2 u_2^{tr} \\ &\quad = u^{rot}(v_1^{tr} - v_2^{tr}) - v^{rot}(u_1^{tr} + u_2^{tr}) \end{aligned} \tag{6}$$

Substituting the respective motions of camera 1 and camera 2 in (2) into (6), we obtain:

$$\begin{aligned} &(v_1 + v_2)U + (u_2 - u_1)V + \frac{(u_1 - u_2)y - (v_1 + v_2)x}{f} W \\ &\quad = 2r(u^{rot}\alpha + v^{rot}\beta) - (u_1 + u_2)\alpha r - (v_1 - v_2)\beta r \end{aligned} \tag{7}$$

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:

$$\frac{u_1 - u_2}{v_1 + v_2} = \frac{Wx - fU}{Wy - fV} \tag{8}$$

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, 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 (7). Collecting all the N equations from the entire set of matching points, we arrive at:

$$\mathbf{A}_{(N\times 3)}\mathbf{x}_1 = -r\mathbf{B}_{(N\times 6)}\mathbf{x}_2 \tag{9}$$

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 ith row of $\mathbf{A}$ and $\mathbf{B}$ are as follows:

$$\begin{aligned} \mathbf{a}_i &= \left[v_1^i + v_2^i,\ u_2^i - u_1^i,\ \frac{(u_1^i - u_2^i)y^i}{f} - \frac{(v_1^i + v_2^i)x^i}{f} \right] \\ \mathbf{b}_i &= \left[u_2^i + u_1^i, v_1^i - v_2^i, \frac{x^{i2} - y^{i2}}{f}, x^i, -y^i, \frac{x^i y^i}{f} \right] \end{aligned} \tag{10}$$

The right-hand-side (RHS) of (7) or 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 (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 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 (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 (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 (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 (Sect. 5.2.2) to solve for the rotation. We then substitute the rotation estimate $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ back into $\mathbf{x}_2$ of (9) and form a new equation:

$$\begin{aligned} &\underbrace{[\mathbf{A}, \mathbf{B}\mathbf{x}_2]} \cdot \underbrace{[\mathbf{x}_1, r]^T} = 0 \\ &\quad \widetilde{\mathbf{A}}_{N\times 4} \cdot \quad \widetilde{\mathbf{x}}_{\mathbf{1}} \quad = 0 \end{aligned} \tag{11}$$

[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!

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 Sect. 5.4.
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 (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 (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.

Assume the singular values of $\widetilde{\mathbf{A}}$ are $\tau_1 \geq \cdots \geq \tau_4 \geq 0$. 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 (11) is homogeneous. 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 (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 (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.

[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.

5.2.2 Stage 2: Recover the Global Rotation

It is not advisable to compute (α, β, γ) from (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 of the two cameras, we obtain a system of equations in the form of:

$$\mathbf{M} \cdot \Big[\underbrace{\alpha, \beta, \gamma}_{\Theta_1}, \underbrace{\frac{r}{s}\beta, \frac{r}{s}\gamma, \frac{r}{s}\alpha\beta, \frac{r}{s}\alpha\gamma, \frac{r}{s}\gamma\beta, \frac{r}{s}(\beta^2 - \gamma^2)}_{\Theta_2 = \frac{r}{s}\Phi}\Big]^T = \mathbf{d} \tag{12}$$

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 $\mathbf{m}_{\text{cam1}}$ and $\mathbf{m}_{\text{cam2}}$, 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:

$$\begin{aligned}
\mathbf{m}_{\text{cam1}} = \Big[& x_1 W - Uf - \frac{x_1 y_1 V - U y_1^2}{f}, \\
& y_1 W - Vf + \frac{x_1 y_1 U - V x_1^2}{f}, \\
& x_1 U + y_1 V - \frac{(x_1^2 + y_1^2) W}{f}, \\
& u_1, v_1, \frac{x_1^2 - y_1^2}{f}, -y_1, x_1, \frac{x_1 y_1}{f} \Big] \\
\mathbf{m}_{\text{cam2}} = \Big[& x_2 W - Uf - \frac{x_2 y_2 V + U y_2^2}{f}, \\
& -y_2 W - Vf - \frac{x_2 y_2 U + V x_2^2}{f}, \\
& x_2 U - y_2 V - \frac{(x_2^2 + y_2^2) W}{f}, \\
& -u_2, v_2, \frac{y_2^2 - x_2^2}{f}, -y_2, -x_2, \frac{x_2 y_2}{f} \Big]
\end{aligned} \tag{13}$$

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

$$\begin{aligned}
\mathbf{d}_{\text{cam1}} &= v_1 U - u_1 V + \frac{W(u_1 y_1 - v_1 x_1)}{f} \\
\mathbf{d}_{\text{cam2}} &= \frac{W(v_2 x_2 - u_2 y_2)}{f} - u_2 V_2 - v_2 U_2
\end{aligned} \tag{14}$$

Due to the existence of six higher order terms in Θ_2, (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 (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:

$$[\,\mathbf{m}_1, \ldots, \mathbf{m}_5\,] \cdot \left[\alpha, \beta, \gamma, \frac{r}{s}\alpha, \frac{r}{s}\beta\right]^T = \mathbf{d} \tag{15}$$

where $\mathbf{m}_i$ is the ith 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 (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 (12) leads to a standard linear system:

$$\left[\mathbf{m}_1,\ \mathbf{m}_2,\ \mathbf{m}_3\ , [\mathbf{m}_4 \ldots \mathbf{m}_9]\hat{\Phi}_0\right] \cdot \left[\alpha, \beta, \gamma, \frac{r}{s}\right]^T = \mathbf{d} \tag{16}$$

Solve the above equation for an updated rotation estimate. If necessary, we substitute the newly obtained estimate back into (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 (12). Algorithm 2 presents our linear rotation estimation algorithm, carried out in three steps.

5.3 Extended Quasi-Parallax for Multiple Camera Pairs

Figure 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 (9). Suppose N_c is the total number of matching points. Collecting the respective equations from all the pairs in the system, we have:

$$\mathbf{A}^*_{(N_c \times 3)}\mathbf{x}_1 = r\mathbf{B}^*_{(N_c \times 9)}\mathbf{x}_3 \tag{17}$$

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 (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 (17) into Algorithm 1 and follow the procedure accordingly. Once the translation is recovered, we then proceed to the rotation estimation stage, described in Algorithm 2. The original system of equations in (12) is slightly modified to:

$$\mathbf{M}^* \cdot \Big[\underbrace{\alpha, \beta, \gamma}_{\Theta_1}, \underbrace{\frac{r}{s}[\alpha, \beta, \gamma, \alpha\beta, \alpha\gamma, \gamma\beta, \alpha^2, \beta^2, \gamma^2]}_{\Theta_3 = \frac{r}{s}\Phi^*}\Big]^T = \mathbf{d}^* \tag{18}$$

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 (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

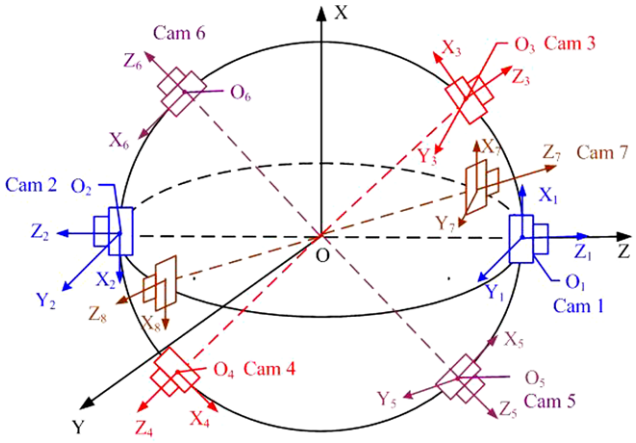


Fig. 4 Multiple camera pairs configured as a compound eye. The global coordinate is $\{O - XYZ\}$ and the kth camera's coordinate is $\{O_k - X_k Y_k Z_k\}$. The system is made up of camera pairs placed at diametrically opposite positions, such as the camera pair 3 and 4

Algorithm 2 Linear Rotation Recovery Algorithm

1: **Linearize:** Directly solve (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 $[\,\mathbf{m}_1, \ldots, \mathbf{m}_5\,]$.
2: **Refine:** Given the estimate $\hat{\omega}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0)^T$, compute $\hat{\Phi}_0 = \Phi|_{\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0}$ and form (16). Solve (16) by linear least squares for a new estimate of ω.
3: **Iterate:** Iterate Step 2, if necessary, until convergence.

Z-axis, introducing more coupling terms between the induced translation and the rotation.

The **Linearize** step in (15) needs to be modified too to incorporate the $\frac{r}{s}\gamma$ term:

$$\left[\mathbf{m}_1^*, \ldots, \mathbf{m}_6^*\right] \cdot \left[\alpha, \beta, \gamma, \frac{r}{s}\alpha, \frac{r}{s}\beta, \frac{r}{s}\gamma\right]^T = \mathbf{d}^* \tag{19}$$

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

5.4 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 (10). The normalization matrix for $\tilde{\mathbf{A}}$ in (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 ith observed vector can be written as:

$$\Delta\mathbf{a}_i = \left[\Delta v_1^i + \Delta v_2^i, \Delta u_2^i - \Delta u_1^i, (\Delta u_1^i - \Delta u_2^i)\frac{y^i}{f} - (\Delta v_1^i + \Delta v_2^i)\frac{x^i}{f}\right] \tag{20}$$

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

$$\mathbf{C}_i = 2\sigma_v^2 \begin{bmatrix} 1 & 0 & -\frac{x^i}{f} \\ 0 & 1 & -\frac{y^i}{f} \\ -\frac{x^i}{f} & -\frac{y^i}{f} & \frac{x^{i2}+y^{i2}}{f^2} \end{bmatrix} \tag{21}$$

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:

$$(\mathbf{A}\mathbf{H})(\mathbf{H}^{-1}\mathbf{x}_1) = 0 \quad \Longrightarrow \quad \bar{\mathbf{A}}\bar{\mathbf{x}}_1 = 0 \tag{22}$$

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.

6 Numerical Characterization

6.1 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 performance 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.

6.2 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

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 $\mathbf{p}_i$ $(i = 1, \ldots, N_c)$ in each 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 $N_1 + N_2 + 6$ variables. There are $N_1 + N_2$ depth unknowns forming the vector $\mathbf{Z}$ and 6 global motion parameters. For the sake of simplicity, we assume $N_1 = N_2$ and r is known.

$$J(\mathbf{v}, \omega, \mathbf{Z}) = \sum_{c=1}^{2} \sum_{i=1}^{N_c} [(\hat{u}_c^i - u_c^i)^2 + (\hat{v}_c^i - v_c^i)^2] \qquad (23)$$

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 (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:

$$\mathbf{F} = \sum_{\mathbf{k} \in D} \left(\sum_{i=1}^{N} \frac{\partial \mathbf{q}_i}{\partial \mathbf{k}}^T \frac{\partial \mathbf{q}_i}{\partial \mathbf{k}} \right) \Bigg|_{\mathbf{k}=(\mathbf{v}, \omega, Z_1, \ldots, Z_N)} \qquad (24)$$

where N is the total number of feature points and $\mathbf{q}_i$ is the optical flow measured at the ith 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 3 m to 7 m in one camera, and from 5 m to 10 m 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 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, 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 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 Sect. 5.2.2 to recover the rotation. Note that since ω_z is not present in our $\mathbf{Q}$, all the associated entries are zero.

Tables 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.

7 Experimental Results

Several experiments were conducted to test our method under different scenarios. The 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

Table 1 Fisher matrices of two methods with ambiguities highlighted, QP refers to Quasi Parallax and BA refers to Bundle adjustment

(a) QP-based with FOV = 50° each

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

(b) BA-based with FOV = 50° 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

(c) QP-based with FOV = 100° 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

(d) BA-based with FOV = 100° 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

of the total magnitude of the translational flow and that of the rotational flow from all feature points. The image resolution used in this paper is 512×512 pixels.

7.1 Experiment on Range Image

This set of experiment uses the Brown range image database (Lee and Huang 2000) which contains many static natural scenes. Figure 5 shows the forest scene we used with depths ranging from 3 m to 10 m. 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 $\varepsilon = 0.2$, 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 nonlinear one. The error of each simulation was computed as the average of the errors from all the trials.

7.1.1 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.[3]

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.

Table 2 In the forest scene, three types of motions corresponding to $\varepsilon = 0.2$, $\varepsilon = 1$, $\varepsilon = 5$ are executed

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

Fig. 5 Range image of a forest scene used. Intensity represents depth with distant object looking brighter. Regions with no range data appears *black*

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

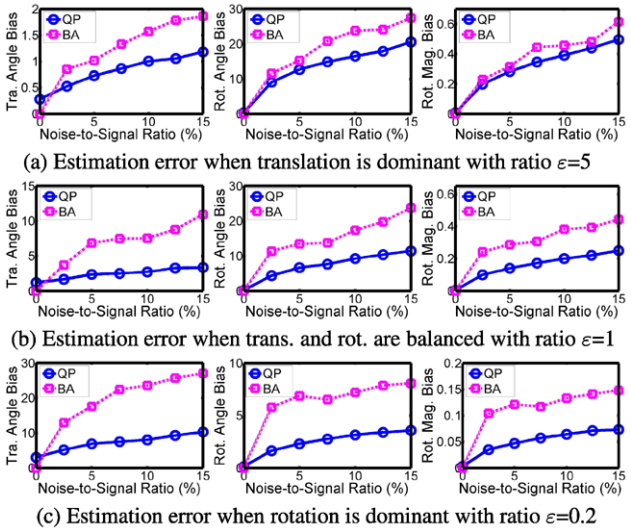


Fig. 6 Motion recovery of the two methods with a pair of cameras with narrow FOV of 15° each. QP stands for our Quasi-parallax method, while BA denotes bundle adjustment. The *three columns*, starting *from the left*, depict the estimation errors in translation direction, rotation direction, and rotation magnitude respectively. Note that the y-axis for different diagrams may have different scales

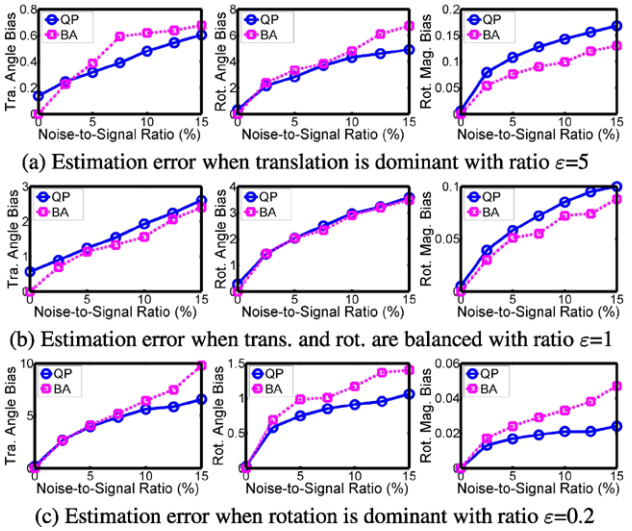


Fig. 7 Motion recovery of the two methods with a pair of laterally-positioned cameras with 50° FOV each

Take the recovery of the translation direction as an example (the left column in Fig. 6). When $\varepsilon = 0.2$, 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 1.86 GHz 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 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 Fig. 6(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.

7.1.2 A Lateral Pair of Avian/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 Fig. 7 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.

We picked 341 matching points for the linear method and a comparable number of 1058 feature points for BA. Under $\varepsilon = 1$ in Fig. 7, 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.

7.1.3 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 Fig. 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 Fig. 8, 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, 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.

7.1.4 Effect of Calibration Errors

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 compound. This is especially pertinent in a biological compound eye

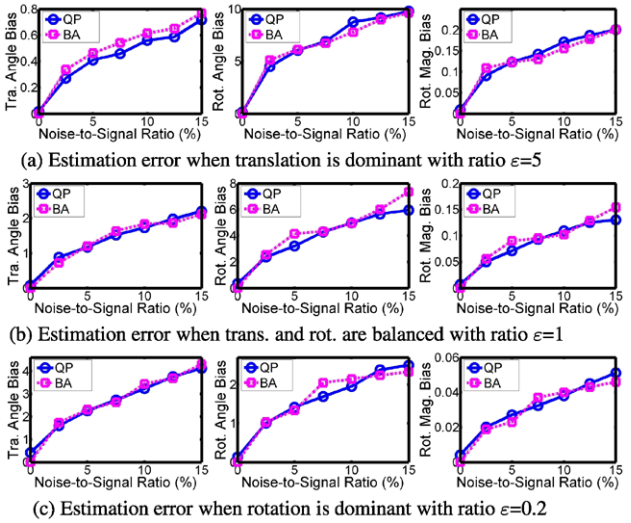


Fig. 8 Motion recovery of the two methods in a compound eye with small number of facets

system, where the head is not a perfect sphere. 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 errors. We used the same compound eye set-up in the previous Sect. 7.1.3 and the same global motions as in Table 2.

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 ith ($i = 1, \ldots, 9$) 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 Fig. 8, our method now gains a distinct advantage over the BA as shown in Fig. 9. 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 the two opposing cameras Cam1 and Cam2 in Fig. 4 do not have

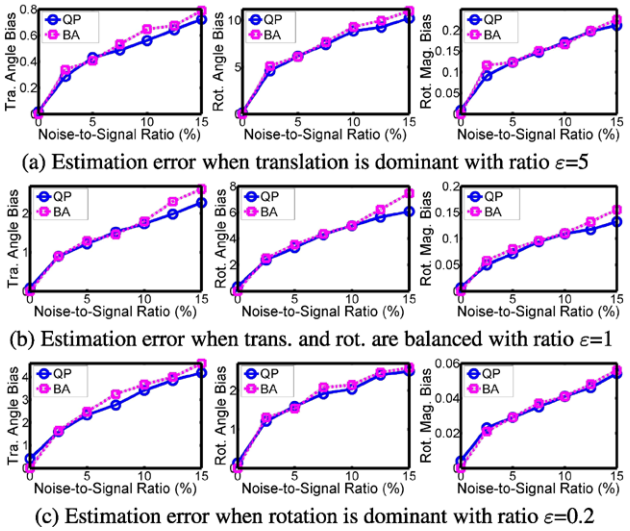


Fig. 9 Motion recovery of the two methods in a compound eye placed on a non-perfect spherical substrate

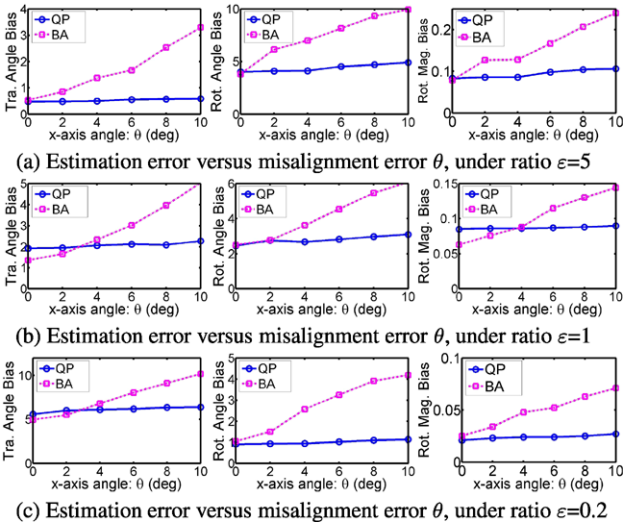


Fig. 10 Motion recovery of the two methods in a compound eye with misalignment error θ around the X-axis

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 Fig. 10 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 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).

7.1.5 Effect of Radius r

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 0.5 m to 3 m. A total of 185 matching points were selected for the linear method while 536 feature points were for bundle adjustment. The flow input of the system was subject to a 10% isotropic noise.

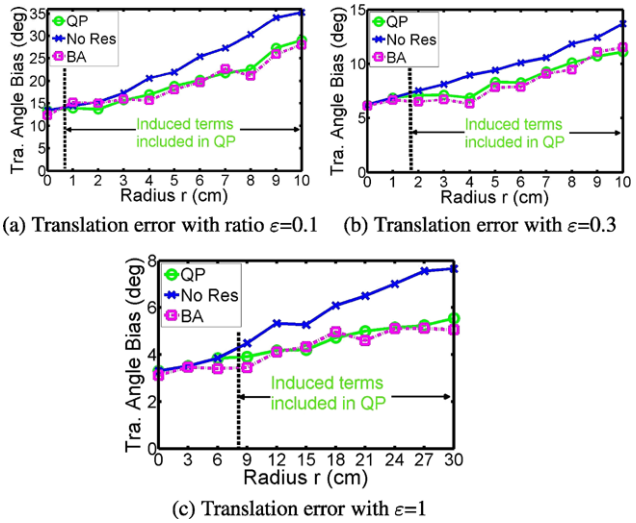
Recall that our formulation modeled these induced terms as $r\mathbf{Bx_2}$ on the right hand side of (9) and iteratively refined their estimates until (11) is sufficiently close to a homogeneous system of equations. The performance of our method is here compared to one which sets r to zero, thereby ignoring the induced terms totally, and is effectively reduced to the antipodal constraint used in the spherical systems of Lim and Barnes (2007, 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 in Table 3.

Table 3 In the indoor scene, three types of motions corresponding to $\varepsilon = 0.1$, $\varepsilon = 0.3$, $\varepsilon = 1$ are executed

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

Referring to Fig. 11, 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 0.09 m, the estimate became worse off by a significant margin if the induced terms were ignored.

The vertical dashed lines in Fig. 11 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.



(a) Translation error with ratio ε=0.1 (b) Translation error with ε=0.3

(c) Translation error with ε=1

Fig. 11 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 (9). The *vertical dashed lines* indicate the radius r above which our algorithm deems the induced terms as significant

Fig. 12 An indoor scene for the real-image experiment

7.2 Experiment on Real Image

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 2 m to 5 m. 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 Fig. 12. We tested three sets of global motions (see Fig. 13 for details), with ε roughly equal to 0.2, 1 and 5 respectively. An average of 140 matching points were selected for the linear method while 850 feature points were selected for the BA. Figure 13 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.

8 Discussion and 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 lateral eye arrangement is realized in many animals including vertebrates such as birds and fishes, and invertebrates such as insects with their compound eyes. Though theories have been

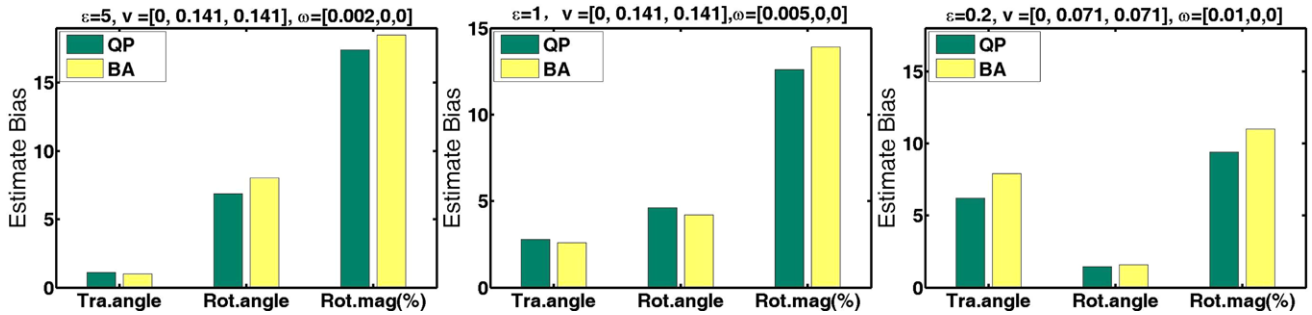


Fig. 13 Motion recovery of the two methods using real image. The unit of translation $\mathbf{v}$ is m/s and the unit of rotation ω is rad/s. Bias plotted in degrees for the two direction estimates and in % for the magnitude estimate

advanced about how these animals exploit this special eye topography to accomplish complex visuomotor tasks with limited brain power, these theories lack computational underpinning.

Our paper 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. Finally, it is 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.

References

Anandan, P., & Irani, M. (2002). Factorization with uncertainty. *International Journal of Computer Vision*, *49*(2–3), 101–116.

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

Blanke, H., Nalbach, H.-O., & Varju, D. (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.

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

Clark, J., & Yuille, A. (1994). *Data fusion for sensory information processing*. Dordrecht: Kluwer Academic.

Coombs, D., & Roberts, K. (1993). Centering behavior using peripheral vision. In *CVPR* (pp. 440–451).

Davies, M. N. O., & Green, P. R. (1994). Multiple sources of depth information: an ecological approach. In *Perception and motor control in birds: an ecological approach* (pp. 339–356). Berlin: Springer.

Franz, M. O. et al. (1998a). Learning view graphs for robot navigation. *Autonomous Robots*, *5*, 111–125.

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

Haag, J., & Borst, A. (2001). Recurrent network interactions underlying flow-field selectivity of visual interneurons. *Journal of Neuroscience*, *21*, 5685–5692.

Hartley, R. (1997). In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(6), 580–593.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Heeger, D. J., & Jepson, A. D. (1992). Subspace methods for recovering rigid motion: I. algorithm and implementation. *International Journal of Computer Vision*, *7*(2), 95–117.

Hornsey, R. et al. (2004). Electronic compound eye image sensor: construction and calibration. In *Proc. SPIE* (pp. 13–24).

Huber, S. A., & Bülthoff, H. H. (1998). Simulation and robot implementation of visual orientation behaviors of flies. In *From animals to animats 5, Proc. of SAB'98* (pp. 77–85).

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, K., Kim, J., & Lee, L. P. (2006). Biologically inspired artificial compound eyes. *Science*, *312*, 557–561.

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, J., Jeong, K., & Lee, L. P. (2005). Artificial ommatidia by self-aligned microlenses and waveguides. *Optical Letter*, *30*, 5–7.

Lee, A. B., & Huang, J. (2000). Brown range image database. http://www.dam.brown.edu/ptg/brid/index.html.

Leedan, Y., & Meer, P. (2000). Heteroscedastic regression in computer vision: Problems with bilinear constraint. *International Journal of Computer Vision*, *37*(2), 127–150.

Leonard, C. S., Simpson, J. I., & Graf, W. (1988). Spatial organization of visual messages of the rabbit's cerebellar flocculus. *Journal of Neurophysiology*, *60*, 2073–2090.

Lim, J., & Barnes, N. (2007). Estimation of the epipole using optical flow at antipodal points. In *OMNIVIS* (pp. 1–6).

Lim, J., & Barnes, N. (2008). Directions of egomotion from antipodal points. In *CVPR*.

Longuet-Higgins, H. C., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London, Series B*, *208*, 385–397.

Neumann, J. et al. (2004). Compound eye sensor for 3D ego-motion estimation. In *IEEE international conference on intelligent robots and automation* (pp. 3712–3717).

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

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

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

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

Stewenius, H., & Astrom, K. (2004). Structure and motion problems for multiple rigidly moving cameras. In *ECCV* (pp. 252–263).

Sturm, P. (2005). Multi-view geometry for general camera models. In *CVPR*, (Vol. 1, pp. 206–212).

Thomas, I., & Simoncelli, E. (1994). *Linear structure from motion* (Technical Report). University of Pennsylvania, IRCS.

Triggs, B. et al. (2000). Bundle adjustment—a modern synthesis. In *Vision algorithms: theory and practice* (pp. 298–375). Berlin: Springer.

Tsao, A. T. et al. (1997). Ego-motion estimation using optical flow fields observed from multiple cameras. In *CVPR*.

Wylie, D. R. W., & Frost, B. J. (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, T., & Cheong, L. F. (2003). Understanding the behavior of SFM algorithms: a geometric approach. *International Journal of Computer Vision*, *51*(2), 113–117.

Zhang, Z. Y. (1995). Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(12), 1222–1227.