Determining Motion Directly from Normal Flows upon the use of a Spherical Eye Platform

Tak-Wai Hui
Department of Mechanical and Automation Eng.
The Chinese University of Hong Kong
twhui1@mae.cuhk.edu.hk

Ronald Chung
Vocational Training Council of Hong Kong
rchung@vtc.edu.hk

Abstract

We address the problem of recovering camera motion from video data, which does not require the establishment of feature correspondences or computation of optical flows but from normal flows directly. We have designed an imaging system that has a wide field of view by fixating a number of cameras together to form an approximate spherical eye. With a substantially widened visual field, we discover that estimating the directions of translation and rotation components of the motion separately are possible and particularly efficient. In addition, the inherent ambiguities between translation and rotation also disappear. Magnitude of rotation is recovered subsequently. Experimental results on synthetic and real image data are provided. The results show that not only the accuracy of motion estimation is comparable to those of the state-of-the-art methods that require explicit feature correspondences or optical flows, but also a faster computation time.

1. Introduction

Determination of the relative motion in space between observer and surrounding environment is important to various applications including augmented reality, 3D reconstruction, and visual control. The translation magnitude of the motion is generally not determinable and left as an overall arbitrary scale relative to object depth because of the well-known ambiguity between object size-depth and translation speed. This paper presents a direct method to determine the five degrees of freedom (DoFs) – the direction of translation and the full rotation of a camera moving in a static scene.

The existing works on motion determination are largely about establishing explicit correspondences across images. The correspondences might be in the form of optical flows (also known as full flows) using monocular camera [14], [26], [33], multiple cameras [17] and spherical camera [24], [25], or correspondences over distinct features using monocular camera [31], [23], [28], multiple cameras [32], [20], [19] and spherical camera [21].

Optical flow induced by the spatial motion at any image position is only partially observable in general due to the familiar aperture problem. The apparent flow, termed the normal flow, which is the component of the optical flow along or opposite to the direction of the local intensity gradient is fully observable. The partial observability of the flow is what makes full flow computation and in turn motion determination a challenge. On the other hand, the feature-based schemes require tracking of distinct feature points, which might not be always present in the video. Ambiguity in establishing correspondences across multiple images exists if the imaged scene contains repetitive patterns. This in turn affects the accuracy in determining camera motion.

A few methods have been proposed to determine camera motion from normal flows directly without ever requiring the full flows to be recovered explicitly. They are collectively known as the direct methods. Such methods arose for natural reasons. Unlike optical flow, normal flow can be obtained directly from image data, without involving minimization of certain cost functional [15], [6], [36] which is often computationally demanding. Even though each normal flow represents only partial information, since the total number of image positions where normal flow is observable generally far exceeds the number of motion parameters, one would expect that motion can be recovered directly from normal flows without imposing additional assumptions like smoothness onto the image flow field. The smoothness assumption is bound to be invalid at many image positions if the imaged scene is non-smooth. In addition, the invention of normal flow measurement camera gives additional support to use normal flow [29].

Standard camera usually has limited field of view (FoV). The video perceived under say a pure translation of the camera in the x-direction would be similar to that under a pure left-hand rotation about the y-axis, where the x and y axes are the orthogonal coordinate axes of the image domain.

In other words, there would be ambiguity in distinguishing motion from the video information. In the biological world, flying insects have wide FoV vision system. They can execute navigational tasks accurately relying on the visual clue from optical flows [35].

In this paper, we present a direct method that uses an approximate spherical eye to recover motion parameters in a static scene. The approximate spherical eye comprises a number of cameras that have optical centers placed near one another without necessary having overlapped FoV. The advantage of having widened visual view is three-fold. First, the translation's direction and the rotation's direction of the motion can be recovered separately in an efficient manner. This avoids error propagation from the recovery of translation to that of rotation, and vice versa. Second, tighter constraints on the motion solution (two pairs of special normal flows are enough to reduce the possible motion solution by $\frac{3}{4}$ of the motion space) are available to improve the result significantly and in turn the computation speed. Third, large visual field could make motion estimation more accurate [16], [10], [11]. This work does not attempt to claim the novelty of advocating the concept that larger visual field incurs better accuracy in motion estimation. The main contribution is instead to provide a mechanism of how motion ambiguity arisen from the use of normal flows could be reduced by separating the translation and rotation motion components through the use of three particular subsets of normal flows. This has not been analyzed before. We do not fuse multiple visual inputs from multiple cameras in a loosely manner, but we consider the underlying geometrical constraints in a coherent and effective way.

2. Related Works

Direct methods determine camera motion from normal flows without prior computation of full flows. Most of the direct methods in the literature use monocular camera with limited FoV. Horn and Weldon required the camera undergoing pure translation, pure rotation, or general motion with known rotation [16]. Aloimonos *et al.* extended the work of Horn and Weldon to general motion by assuming a bounded rotational magnitude [1]. Fermüller *et al.* proposed to select special image points that form some global patterns [8]. The boundary of each pattern is generally difficult to extract due to the sparse normal flow field. Silva *et al.* proposed several algorithms which search lines and curves to estimate the motion parameters [34]. Yet only a limited number of normal flows participated in recovering egomotion. Brodský *et al.* proposed to use minimization of variation of local depth [5]. The scene in view is assumed to be piecewise smooth. We proposed a two-stage iterative method which requires to work in a high dimensional space [18].

There are a few direct methods in the literature that use wide FoV. Nelson *et al.* turned the recovery of camera motion into three sub-problems, each involving one rotational and two translational parameters (translation and rotation are mixed up) [30]. This partial separation of motion components is only possible when normal flows are located at the three equators which are perpendicular to the three principal axes of the camera's coordinate frame respectively. Fermüller *et al.* also presented the use of global patterns [8] for the case of spherical eye [7]. Baker *et al.* extended their previous work from planar eye [5] to spherical eye [2]. The scene in view is still assumed to be piecewise smooth. Multiple solutions from individual cameras are fused to reduce the ambiguity of the solution.

Brodský *et al.* provided an analysis about the conditions when apparent flows become ambiguous [4]. Fermüller and Aloimonos characterized the structure of rigid motion fields [9], ambiguity in structure from motion using planar and spherical eyes [10], and also the observability of 3D motion under different fields of view [11].

Our proposed method, being a direct one, provides an alternative approach to recover camera motion without the need of matching feature correspondences and recovery of full optical flows as the current state-of-the-art methods. Our work is related to [24] that both of us determine the directions of translation and rotation from general motion using merely the direction component of flow vectors. Unlike their work, we utilize normal flows which are directly observable instead of prior computation of full flows. Our algorithm is developed for a multi-camera rig but not for a perfectly spherical eye. Moreover, we do not demand each pair of flow vectors located at opposite image positions on the image sphere. Unlike the the rotation independent constraint in [21], we do not require prior knowledge about scene depth to determine translation. In contrast to the work of [30], the separation of motion components is not limited to image positions at the equator of each principal axis of the image sphere. Unlike the works of [2], [17], [25], our strategy is to separate the directions of translation and rotation from general motion. This avoids the problem of error propagation from translation to rotation, and vice versa. In contrast to [2], we use visual clues from all cameras simultaneously and without assuming piecewise-smooth scene.

3. Background

Consider a camera undergoes a general motion with an instantaneous translational velocity $\mathbf{t}$ (a 3-vector representing the translation direction and magnitude) and instantaneous angular velocity $\mathbf{w}$ (a 3-vector representing the rotation axis and magnitude) in a stationary environment, as depicted in Figure 1a. Let the camera's focal length be f. Suppose all spatial quantities are with reference to the camera-centered coordinate system C (X_C-Y_C-Z_C). The instantaneous velocity of any 3D object point $\mathbf{X} = (X, Y, Z)^T$ of the

scene relative to frame C is:

$$\dot{\mathbf{X}} = -\mathbf{t} - \mathbf{w} \times \mathbf{X}. \tag{1}$$

The 3D point $\mathbf{X}$ projects under perspective projection to the camera's image plane at the image position $\mathbf{x} = (x, y)^T$ with respect to the origin O of the image plane. The 3D point $\mathbf{X}$ and the image position $\mathbf{x}$ are related by:

$$\mathbf{X} \cong \tilde{\mathbf{x}}, \tag{2}$$

where $\cong$ denotes equality up to arbitrary nonzero scale, and $\tilde{\mathbf{x}} = (x/f, y/f, 1)^T$ represents the projective coordinates of $\mathbf{x}$. The optical flow $\dot{\mathbf{x}}$ at the image position $\mathbf{x}$ is given by:

$$\dot{\mathbf{x}} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \end{bmatrix} \left(f(\mathbf{k} \times (\tilde{\mathbf{x}} \times \mathbf{t})) / Z + f(\mathbf{k} \times ((\tilde{\mathbf{x}} \times \mathbf{w}) \times \tilde{\mathbf{x}})) \right), \tag{3}$$

where $\mathbf{k} = (0, 0, 1)^T$ is a unit vector in the direction of the optical axis of the camera and the matrix $\begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \end{bmatrix}$ transforms optical flow from 3-vector (in projective coordinates relative to the camera center) to 2-vector (relative to the origin of image plane).

If the image plane is warped to a spherical imaging surface with focal length f (as shown in Figure 1b), image position becomes $\mathbf{x}^s = f\mathbf{X}/||\mathbf{X}||$. The optical flow $\dot{\mathbf{x}}^s$ which is tangential to the imaging spherical surface at $\mathbf{x}^s$ is given by:

$$\dot{\mathbf{x}}^s = (\mathbf{x}^s \times (\mathbf{x}^s \times \mathbf{t})) / \|\mathbf{X}\| f + \mathbf{x}^s \times \mathbf{w}. \tag{4}$$

It can be seen from (4) that the magnitude $||\mathbf{t}||$ of translation cannot be decoupled from the scene depth $||\mathbf{X}||$ from visual information alone (in the sense that a closer and smaller object observed under a slower translation could appear the same in the video data as a farther and bigger object observed under a faster translation), yet the rotational component of the motion is independent of scene depth.

In general, optical flow at any image position is not directly observable from the image because of the well-known aperture problem. Only the projected component of the flow to the spatial intensity gradient $\mathbf{n}$ (normalized to a unit vector) at the position, in the name of normal flow, is directly observable. By using the Brightness Constancy Constraint Equation (BCCE) [15], we can relate optical flow $\dot{\mathbf{x}}$ and spatial intensity gradient ∇I in planar image as:

$$\nabla I \cdot \dot{\mathbf{x}} + I_t = 0. \tag{5}$$

Normal flow can be expressed as:

$$\dot{\mathbf{x}}_n = (\dot{\mathbf{x}} \cdot \mathbf{n})\, \mathbf{n} = -I_t \nabla I / \|\nabla I\|^2, \tag{6}$$

where ∇I and I_t denote the spatial gradient and temporal gradient of the video data $I(\mathbf{x}, t)$ at frame t respectively. Similar expressions for normal flows induced on the spherical imaging surface can also be derived.

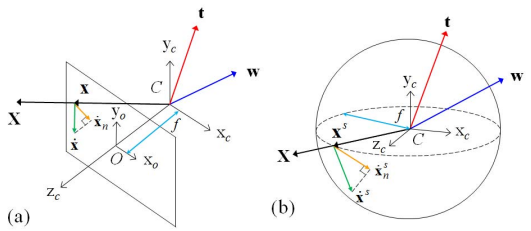

Figure 1. Image formation on (a) an image plane and (b) an image sphere. The camera moves with a translation $\mathbf{t}$ and a rotation $\mathbf{w}$. A 3D scene point $\mathbf{X}$ projects onto the image plane at $\mathbf{x}$ and induces optical flow $\dot{\mathbf{x}}$ and normal flow $\dot{\mathbf{x}}_n$ there. Similarly, optical flow $\dot{\mathbf{x}}^s$ and normal flow $\dot{\mathbf{x}}^s_n$ are induced at $\mathbf{x}^s$ on the spherical surface.

4. Approximation of Spherical Eye by Multiple Cameras

There is practical difficulty in realizing an ideal spherical eye. One way of constructing a spherical eye is to stitch two omnidirectional cameras together in a back-to-back configuration. However, the acquired image has non-uniform resolution over the image domain. In this work, we explore the use of multiple standard cameras to approximate the spherical eye. We bundle the cameras together with their optical centers close to one another and with their visual fields distinct. If the optical centers of the cameras can be made exactly concurrent, the multi-camera system mimics a spherical eye.

In practice it is not possible to bundle the cameras as exactly, and there are bound to be certain nonzero baseline distances between the cameras. There is evidence indicating that even with such an imperfect imaging system, the gain (in having a wider FoV) generally outweighs the loss (from the non-concurrency of the optical centers), and substantial improvement in recovering motion is possible [19].

With a number of cameras stitched together, we seek to recover the 5 DoFs of the rigid motion of the camera rig directly from the normal flows observed in the various cameras. Below we first outline how normal flows in the multiple cameras are related to the camera rig motion when the baseline distances between the cameras are small compared with the overall scene depth from the camera rig.

Suppose the image point $\mathbf{x}_i$ in the i^{th} camera is projected by a 3D point $\mathbf{X}_i$ with depth Z_i (with respect to the camera coordinate frame C_i), and at the image position $\mathbf{x}_i$ the local camera motion $(\mathbf{t}_i, \mathbf{w}_i)$ induces a full flow $\dot{\mathbf{x}}_i$. By projecting the full flow $\dot{\mathbf{x}}_i$ to the spatial gradient $\mathbf{n}_i$ (a unit vector) at the image position $\mathbf{x}_i$, we have the normal flow magnitude:

$$\begin{aligned} |\dot{\mathbf{x}}_i \cdot \mathbf{n}_i| = \Big| & f_i \mathbf{t}_i \cdot ((\mathbf{n}_i \times \mathbf{k}_i) \times \tilde{\mathbf{x}}_i) / Z_i \\ & + f_i \mathbf{w}_i \cdot (\tilde{\mathbf{x}}_i \times ((\mathbf{n}_i \times \mathbf{k}_i) \times \tilde{\mathbf{x}}_i)) \Big|. \end{aligned} \tag{7}$$

We can then relate the rigid motion $(\mathbf{t}, \mathbf{w})$ of the camera

rig in the global coordinate system C_0 (which we take as the center of the camera rig) to the motion ($\mathbf{t}_i$, $\mathbf{w}_i$) of the i^{th} camera in its own local coordinate system C_i, as:

$$\mathbf{t}_i = \mathbf{R}_i^T \left(\mathbf{w} \times \mathbf{b}_i + \mathbf{t} \right), \tag{8}$$

$$\mathbf{w}_i = \mathbf{R}_i^T \mathbf{w}, \tag{9}$$

where $\mathbf{b}_i$ and $\mathbf{R}_i$ are the position (baseline) and rotation of the i^{th} camera with respect to the global frame C_0.

By a few algebraic manipulations over (6), (7), (8), and (9), normal flow $\dot{\mathbf{x}}_{ni}$ in the i^{th} camera can be related to the desired motion parameters $\mathbf{t}$ and $\mathbf{w}$ by:

$$d_i = \frac{\|\dot{\mathbf{x}}_{ni}\|}{f_i} = -\frac{1}{Z_i}\mathbf{t}\cdot(\mathbf{R}_i\mathbf{a}_{ti}) + \mathbf{w}\cdot\left(\mathbf{R}_i\mathbf{a}_{wi} - \frac{1}{Z_i}\mathbf{b}_i \times \mathbf{R}_i\mathbf{a}_{ti}\right) \tag{10}$$

where

$$\mathbf{a}_{ti} = \tilde{\mathbf{x}}_i \times (\sin\theta_i, -\cos\theta_i, 0)^T, \tag{11}$$

$$\mathbf{a}_{wi} = \mathbf{a}_{ti} \times \tilde{\mathbf{x}}_i, \tag{12}$$

are terms related to the normal flow with orientation θ_i at the image position $\mathbf{x}_i$ of the i^{th} camera. Equation (10) extends a variant of the brightness change equation [16] from the case of single camera [18] to the case of multiple cameras.

The two vector entities $\mathbf{a}_{ti}$, $\mathbf{a}_{wi}$, and the image position vector $\tilde{\mathbf{x}}_i$ are mutually perpendicular to one another. The distribution of $\mathbf{a}_{ti}$ lies on the surface of a cylinder in 3-space. In particular, $\mathbf{a}_{ti}$ is orthogonal to $(\sin\theta_i, -\cos\theta_i, 0)^T$ which is the direction vector of the normal flow $(\dot{\mathbf{x}}_{ni}, 0)^T$ (in projective coordinates) rotated about the camera's optical axis by 90°. This means that $\mathbf{a}_{ti}$, the image position vector $\tilde{\mathbf{x}}_i$, and the normal flow $(\dot{\mathbf{x}}_{ni}, 0)^T$ (in projective coordinates) lie on the plane (Π_i), and $\mathbf{a}_{ti}$ points in the direction governed by (11). While $\mathbf{a}_{wi}$ is orthogonal to that plane and points in the direction governed by (12). This important observation is used to derive the proposed direction constraint in section 5.

In practice, the norm of the baseline vector $\mathbf{b}$ is small compared with the scene depth Z. Moreover, $\|\mathbf{a}_w\| = \|\mathbf{a}_t\|\,\|\tilde{\mathbf{x}}\| \geq \|\mathbf{a}_t\|$. Equation (10) can thus well be approximated to:

$$d_i \approx -\rho_i \hat{\mathbf{t}} \cdot (\mathbf{R}_i\mathbf{a}_{ti}) + \mathbf{w} \cdot (\mathbf{R}_i\mathbf{a}_{wi}), \tag{13}$$

where ρ_i is the $\|\mathbf{t}\|$-scaled inverse scene depth and $\hat{\mathbf{t}}$ is the unit vector of $\mathbf{t}$. The two terms $\mathbf{R}_i\mathbf{a}_{ti}$ and $\mathbf{R}_i\mathbf{a}_{wi}$ indicate that every normal flow data point from each camera is transformed from its local camera system C_i to the global coordinate system C_0 through the rotation matrix $\mathbf{R}_i$. Equation (13) reveals how distinct cameras work together as a single imaging system. Dropping the camera index for simplicity of notations, and combining terms, we express (13) as:

$$d = -\rho\hat{\mathbf{t}} \cdot \mathbf{a}'_t + \mathbf{w} \cdot \mathbf{a}'_w. \tag{14}$$

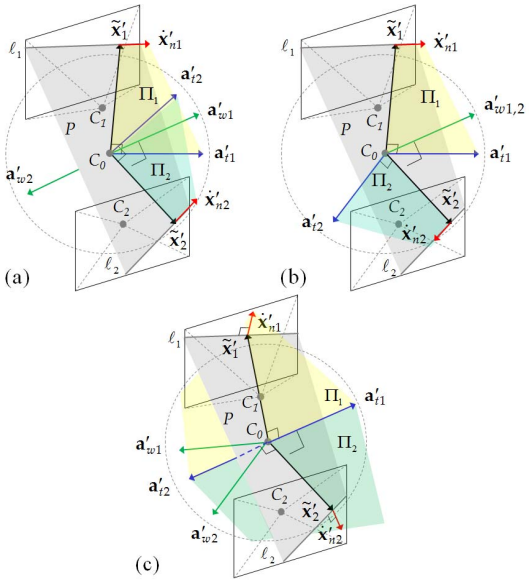

Figure 2. A pair of normal flows $(\dot{\mathbf{x}}'_{n1}, \dot{\mathbf{x}}'_{n2})$ at the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$ is classified as (a) an α-vector pair, (b) a β-vector pair, and (c) a γ-vector pair, if they meet certain conditions.

5. Recovery of Motion Parameters

In our recent work about a direct method [18], a 4D search method was proposed to recover both directions of translation and rotation. Our experimental results show that for some scenes, the constraint could only restrict the motion parameters to a rather large set of possible solutions. This work contributes two improvements. One is to use a wide FoV imaging system to reduce the ambiguity in motion estimation. In addition, we separate the translation and rotation components in the motion recovery process, and treat them one by one. This makes motion estimation far simpler.

5.1. Classification of a Pair of Normal Flows

Consider the spherical imaging surface approximated by the images planes of several standard cameras (with camera centers possibly mildly non-concurrent). Here, we just use two cameras with centers C_1 and C_2 to illustrate the classification of normal flow pairs in Figure 2. Suppose we have two observable normal flows $\dot{\mathbf{x}}_{n1}$ and $\dot{\mathbf{x}}_{n2}$ at the image positions $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ (3-vectors expressed in projective coordinates) with respect to their local camera coordinates respectively. For the spherical approximation of multiple cameras, the two camera centers C_1 and C_2 coincide to the camera rig's center C_0. All the measured entities are transformed from their local camera coordinates to the camera rig's coordinate system (i.e. $\mathbf{a}'_{ti} = \mathbf{R}_i\mathbf{a}_{ti}$, $\mathbf{a}'_{wi} = \mathbf{R}_i\mathbf{a}_{wi}$, $\tilde{\mathbf{x}}'_i = \mathbf{R}_i\tilde{\mathbf{x}}_i$, $\left(\dot{\mathbf{x}}'_{ni}, 0\right)^T = \mathbf{R}_i\left(\dot{\mathbf{x}}_{ni}, 0\right)^T$).

The two position vectors $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$ span a plane P in 3-space. The two image planes intersect the plane P in lines

l_1 and l_2 respectively. We classify a pair of normal flows $\left(\dot{\mathbf{x}}'_{n1}, \dot{\mathbf{x}}'_{n2}\right)$ into three groups according to their orientations with respect to the lines l_1 and l_2. If $\dot{\mathbf{x}}'_{ni}$ is parallel to l_i and $\dot{\mathbf{x}}_{n1} \cdot \dot{\mathbf{x}}_{n2} > 0$, then we classify the normal flow pair as an α-vector pair (Figure 2a). If $\dot{\mathbf{x}}'_{ni}$ is parallel to l_i and $\dot{\mathbf{x}}_{n1} \cdot \dot{\mathbf{x}}_{n2} < 0$, then we classify the vector pair as a β-vector pair (Figure 2b). If $\dot{\mathbf{x}}'_{ni}$ is orthogonal to l_i and $\dot{\mathbf{x}}_{n1} \cdot \dot{\mathbf{x}}_{n2} < 0$, then the vector pair is defined as a γ-vector pair (Figure 2c).

From the observations concluded in section 4, the vector entity $\mathbf{a}'_{ti}$, the projective image coordinates $\tilde{\mathbf{x}}'_i$, and the normal flow $\left(\dot{\mathbf{x}}'_{ni}, 0\right)^T$ (in projective coordinates) lie on the plane Π_i. This means that $\mathbf{a}'_{t1}$ and $\mathbf{a}'_{t2}$ lie on P (i.e. Π_1, Π_2, and P are coplanar) when the normal flows are α- or β-pair. They are orthogonal to P (i.e. Π_1, Π_2 are orthogonal to P) when the normal flows are γ-pair. The vector entity $\mathbf{a}'_{wi}$ is always orthogonal to the associated Π_i.

5.2. Apparent Flow Separation (AFS) Constraint

Suppose we have an α-vector pair at the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$ of the approximate spherical eye as shown in Figure 2a. Since the planes Π_1, Π_2, and P are coplanar, $\mathbf{a}'_{t1}$ and $\mathbf{a}'_{t2}$ lie on P, and $\mathbf{a}'_{w1}$ and $\mathbf{a}'_{w2}$ are orthogonal to P. According to (12), $\mathbf{a}'_{w1}$ and $\mathbf{a}'_{w2}$ point in opposite directions. Their vector sum can be eliminated if each $\mathbf{a}'_{wi}$ is normalized. Summing the two $\|\mathbf{a}'_w\|$-normalized Equations (14) over the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$, we can reduce the general motion to the case of pure translation, as:

$$\text{AFS}_\alpha(\hat{\mathbf{t}}; \mathbf{x}, \theta) : \hat{\mathbf{t}} \cdot \hat{\mathbf{a}}'_{t1} < 0 \text{ or } \hat{\mathbf{t}} \cdot \hat{\mathbf{a}}'_{t2} < 0, \qquad (15)$$

where $\hat{\mathbf{t}}$ and $\hat{\mathbf{a}}'_{ti}$ are unit vectors of $\mathbf{t}$ and $\mathbf{a}'_{ti}$ respectively for i = 1, 2. The locus is in the form of a compound linear inequality that binds the direction of $\mathbf{t}$.

Suppose we have a β-vector pair. Similar to the derivation for the case of α-vector pair, $\mathbf{a}'_{w1}$ and $\mathbf{a}'_{w2}$ are orthogonal to P and point in the same direction as shown in Figure 2b. Their vector subtraction can be eliminated if each $\mathbf{a}'_{wi}$ is normalized. Subtracting the two $\|\mathbf{a}'_w\|$-normalized Equations (14) at the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$, it can be shown that the rotation component in (14) can be eliminated. We can reach the following:

$$\text{AFS}_\beta(\hat{\mathbf{t}}; \mathbf{x}, \theta, d) : \hat{\mathbf{t}} \cdot \lambda\hat{\mathbf{a}}'_{t1} < 0 \text{ or } \hat{\mathbf{t}} \cdot \lambda\hat{\mathbf{a}}'_{t2} > 0, \qquad (16)$$

where $\lambda = \text{sign}\left(d_1 / \|\mathbf{a}'_{w1}\| - d_2 / \|\mathbf{a}'_{w2}\|\right)$. The locus is again in the form of a compound linear inequality that binds the direction of $\mathbf{t}$.

Consider we have a γ-vector pair at the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$ of the approximate spherical eye. The planes Π_1 and Π_2 are orthogonal to P as shown in Figure 2c. In addition, $\mathbf{a}'_{w1}$ and $\mathbf{a}'_{w2}$ lie on P. According to (12), $\mathbf{a}'_{t1}$ and $\mathbf{a}'_{t2}$ point in opposite directions. Their vector sum can be eliminated if each $\mathbf{a}'_{ti}$ is normalized. Summing the two $\|\mathbf{a}'_t\|$-normalized Equations (14) at the image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$, it can be shown that the translation component in (14) can be eliminated. We can reach the following:

$$\text{AFS}_\gamma(\hat{\mathbf{w}}; \mathbf{x}, \theta) : \hat{\mathbf{w}} \cdot \hat{\mathbf{a}}'_{w1} > 0 \text{ or } \hat{\mathbf{w}} \cdot \hat{\mathbf{a}}'_{w2} > 0, \qquad (17)$$

where $\hat{\mathbf{w}}$ and $\hat{\mathbf{a}}'_{wi}$ are unit vectors of $\mathbf{w}$ and $\mathbf{a}'_{wi}$ respectively for i = 1, 2. In essence, $\text{AFS}_\gamma(\hat{\mathbf{w}}; \mathbf{x}, \theta)$ is a compound linear inequality on the direction of $\mathbf{w}$.

The above AFS constraints, with the exception of the $\text{AFS}_\beta(\hat{\mathbf{t}}; \mathbf{x}, \theta, d)$ constraint, depend only on the direction of the normal flows. A proper threshold value can be used to determine the sign of the scalar λ. All the constraints are in the form of system of linear inequalities, which we can solve them efficiently using Hough-like voting.

Each pair of normal flows that belong to either the α-, β-, or γ-group can trim away the associated motion space (α- and β-vectors for $\hat{\mathbf{t}}$-space, γ-vector for $\hat{\mathbf{w}}$-space) up to half of the motion space, i.e. three-quarter of the 4-D parameter space $\left(\hat{\mathbf{t}}, \hat{\mathbf{w}}\right)$ when combined together. The amount of solution trimming depends upon the angle subtended by the two constraint vectors (i.e.,$\left(-\hat{\mathbf{a}}'_{t1}, -\hat{\mathbf{a}}'_{t2}\right)$ in AFS_α, $\left(-\lambda\hat{\mathbf{a}}'_{t1}, \lambda\hat{\mathbf{a}}'_{t2}\right)$ in AFS_β, and $\left(\hat{\mathbf{a}}'_{w1}, \hat{\mathbf{a}}'_{w2}\right)$ in AFS_γ). Indeed, this angle depends on the angle θ_x which is subtended by the two associated image position $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$. If we define the two constraint vectors as $\mathbf{v}_1$, $\mathbf{v}_2$, and their angular separation as θ_v, it is possible to show that:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \cos\theta_v = -\cos\theta_x, \qquad (18)$$

i.e., $\theta_v = \pi - \theta_x$. This means the wider is the angular separation between the two image positions $\tilde{\mathbf{x}}'_1$ and $\tilde{\mathbf{x}}'_2$, the larger would be the trimming. We call the above constraints as the Apparent Flow Separation (AFS) constraints.

5.3. Apparent Flow Magnitude (AFM) Constraint

The AFS constraints return two probability maps that indicate the the direction of translation and the direction of rotation separately. The solution set can be further refined by going through a second stage which involves normal flow's magnitude by the use of partial detranslation and complete derotation similar to that described in [8]. Here, we combine both partial detranslation and complete derotation techniques together. The magnitude component of normal flows could be more erroneous than its direction component because the temporal resolution of a video is generally lower than its spatial resolution. The approach that uses AFS and AFM in a two-stage manner makes the solution more robust to noise.

In partial detranslation, we need to search for the vector set $V = \{\mathbf{a}_t : \hat{\mathbf{t}} \cdot \mathbf{a}_t = 0\}$ for each $\hat{\mathbf{t}}$. A reduced set of hypothesized $\hat{\mathbf{t}}$ can be generated from the AFS constraint beforehand. An optimal full $\mathbf{w}$ is returned as the least-square-error (LSE) solution of the system of linear equations:

$$\left[\mathbf{a}_w^T\right] \mathbf{w} = [d], \qquad (19)$$

for each hypothesized $\hat{\mathbf{t}}$. A score s_{DeT} which can be given to each $\hat{\mathbf{t}}$ depends on the median of square residues. The smaller is the median value, the higher would be the score. A solution $\hat{\mathbf{w}}$ from (19) which is outside the spherical convex hull of $\{\hat{\mathbf{w}}\}$ resulted from AFS$_\gamma$ is considered to be a degenerate solution. Such a $\hat{\mathbf{t}}$ and also $\hat{\mathbf{w}}$ are removed from the solution set.

In complete derotation, we need to remove rotational component in normal flows by using $\{\mathbf{w}\}$ which is obtained from partial detranslation. The $\mathbf{w}$ which fulfils better the system of linear inequality:

$$\left[\sigma \hat{\mathbf{a}}_t^T\right] \hat{\mathbf{t}} < 0, \tag{20}$$

where $\sigma = \text{sign}\,(d - \mathbf{w} \cdot \mathbf{a}_w)$, will have a higher score s_{DeW} which equals the number of satisfied inequalities. In particular, only those data points that have non-zero $\hat{\mathbf{t}} \cdot \mathbf{a}_t$ are used. Our complete derotation differs from [8] in the way that we express the constraint in terms of a system of linear inequalities which is more computationally efficient. We do not require to generate a set of derotated normal flows, and to consider whether they are constrained to lie in half-planes that are related to the hypothesized focus of expansion (or focus of contraction).

The requirements of $\mathbf{a}_t$ for the partial detranslation and for complete derotation are actually complementary to each other. One requires $\mathbf{a}_t$ orthogonal to $\hat{\mathbf{t}}$ and the other one requires $\mathbf{a}_t$ not orthogonal to $\hat{\mathbf{t}}$. Therefore, we use a scoring scheme such that the score of a candidate of $\hat{\mathbf{t}}$ is the weighted sum of s_{DeT} and s_{DeW}. The weights are the number of $\mathbf{a}_t$ satisfying $\hat{\mathbf{t}} \cdot \mathbf{a}_t = 0$ and $\hat{\mathbf{t}} \cdot \mathbf{a}_t \neq 0$ respectively. The solution for $\hat{\mathbf{t}}$ will be the one having the highest score. The optimal full $\mathbf{w}$ will be the associated $\mathbf{w}$ for that $\hat{\mathbf{t}}$ from the partial detranslation.

6. Experiments and Results

6.1. Simulation

We evaluated the performance of our proposed method (**AFS+AFM**) and a 5-point method (**spher-5-pt RANSAC**) [19]. Both works require to use an approximate spherical camera. For **spher-5-pt RANSAC**, we skipped 3 image frames for every motion estimate as the result was not stable if the baseline was too short. The simulation environment was conducted as close to the real situation as possible. The spherical imaging system consisted of 4 cameras which were positioned in a cross-shaped configuration as shown in Figure 4b. Each camera had a resolution 640 × 480, and focal length was set to 350 pixels. A virtual scene having depth randomly varied from 75*cm* to 125*cm* was used. Each camera in the rig was placed 2*cm* away from the global coordinate frame C_0. All the cameras lay on the same plane nominally. In order to simulate errors introduced from cameras' placement discrepancy, a position vector with random

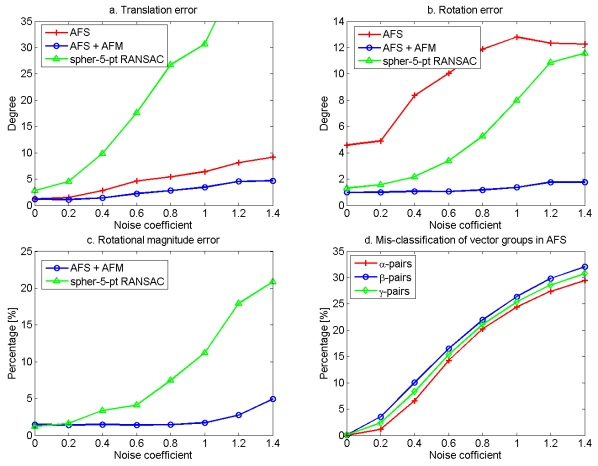

Figure 3. Performance of our proposed method under different levels of Gaussian noise.

orientation and magnitude equals to 1*mm* was added to each baseline vector $\mathbf{b}_i$. A random rotation matrix which was equivalent to a rotation of 1.5° magnitude about a random axial direction was also multiplied to each rotation matrix $\mathbf{R}_i$. To simulate a sparse flow field, we randomly chose only 5% of the flow vectors. To simulate flow extraction error, full flows were corrupted by Gaussian noise. The standard deviation of the noise equals to the noise coefficient times the median of full flows' magnitudes. The camera rig underwent general motion that included both translation and rotation motions randomly generated over all possible directions. The magnitudes of translation and rotation were fixed to 6.67*mm/frame* and 0.4°/*frame* respectively. Results were averaged over 100 trials for each level of noise.

We randomly picked up 2000 pairs of normal flows for each of the α- and β-groups, and another 4000 pairs of normal flows for the γ-group. This means that the estimation of the directions of translation and rotation used the same amount of normal flows. The angular separation θ_x between the two image positions $\tilde{\mathbf{x}}_1'$ and $\tilde{\mathbf{x}}_2'$ in the pair was required to be greater than 150°. We conducted voting for AFS in a coarse-to-fine manner on a unit sphere which was uniformly sampled by the icosahedral best packing [12]. The sampling resolutions were increased from 19.32° to 1.181°. Candidates which reached 98% of the maximum votes or above were considered to be the solutions. The final single solution was defined as the vote-weighted average of this set of candidates.

Figure 3 shows the simulation results. Our method **AFS+AFM** achieved better accuracy in estimating the direction of translation than **spher-5-pt RANSAC** even we just used the AFS constraint. For the recovery of rotation, we achieved a very low estimation error when AFS and AFM constraints are used together. In particular, the angular errors are still within acceptable level even if there exists

more than 30% outliers (noise coefficient = 1.4), 5.183° for translation and 1.764° for rotation with 4.917% magnitude error. The overall result shows that the proposed method has acceptable level of error and is robust against noise even the system is not a perfect spherical eye.

6.2. Real Video

Real image sequences (resolution 640 × 360) were captured using a custom-made imaging system (Fig. 4a) that consists of four Microsoft HD-5000 cameras positioned in a cross-shaped configuration (Fig. 4b). The cameras were intrinsically and extrinsically calibrated using the calibration toolbox [3] and the grid calibration [2] respectively. The baseline vector $\mathbf{b}_i$ for each camera was about 3*cm* from the rig center C_0. The camera system was placed on a computer-controlled xy-table with a manually tunable rotation stage (Fig. 4c). This means that the camera system can be translated in x_0z_0-plane and rotated along y_0-axis (perpendicular to the ground). The method is not limited to planar motion, but for practical reasons, the experiment was carried on the ground plane. Fig. 4d shows a set of images captured by the system during the experiment. The motion ground truth was: $\mathbf{t} = (-0.1736, 0, 0.9848)^T \times 1.5mm/frame$ and $\mathbf{w} = (0, -1, 0)^T \times 0.15°/frame$. Normal flows were calculated by (6). To ease the differentiation process, each image frame was smoothed by 2D Gaussian filter. The spatial and temporal derivatives were calculated using the second order approximation (stencil $\frac{1}{12}$[1 -8 0 8 -1]). Same amount of normal flows was used as the simulation.

We compared the performance of our method against the the state-of-the-art algorithms:

1. **8-pt RANSAC** – A 8-point algorithm utilizes feature correspondences in a RANSAC framework [13].

2. **spher-5-pt RANSAC** – A 5-point algorithm utilizes feature correspondences in RANSAC to estimate motion from an approximate spherical camera [19].

3. **TV-L_1-NL+LM** – A linear method utilizes optical flows to estimate camera motion [33]. Optical flow field was recovered using a TV-L_1 energy minimization with weighted median filtering [36].

4. **TV-L_1-NL+LQP** – A linear quasi-parallax method [17] uses optical flows [36] from pairs of anti-parallel visual rays.

5. **AFD+AFM** – A two-stage direct method utilizes normal flows to estimate motion of monocular camera [18]. Instead of non-uniform sampling of the motion space, we used uniform sampling.

We used Scale Invariant Feature Transform (SIFT) [27] for features detection in **8-pt RANSAC** and **spher-5-pt RANSAC** due to its outstanding performance and availability of the code. We also used the Matlab functions from [22]. In the evaluation, all programs were written in MatLab and they ran on a WinXP PC with 3GHz Pentium D CPU and 2GB RAM. It should be noticed that SIFT was written in C++ and then complied to form a MEX-function for Matlab. The comparison in computation speed is indeed biased to the methods using SIFT.

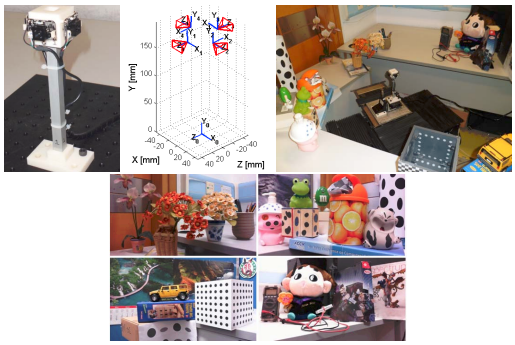

Figure 4. (a) An approximate spherical eye constructed from 4 cameras. (b) Schematic of the vision system. (c) Experimental setup. (d) An image set obtained from the vision system.

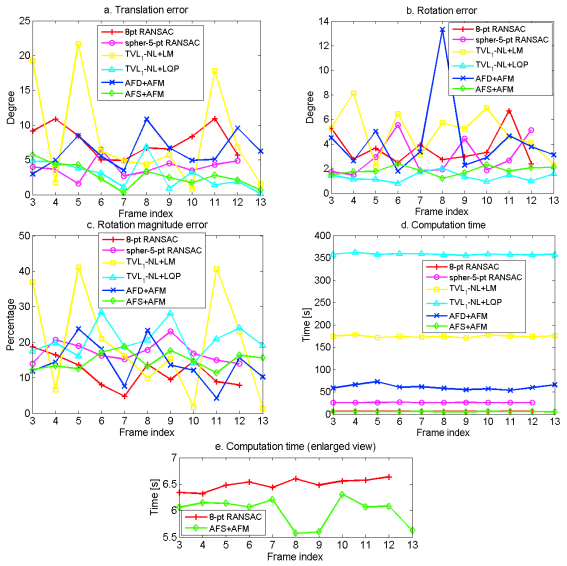

Figure 5. Results for the real-image experiment.

Figure 5 shows the results of the experiment. Only the motion estimations from frame 3 to frame 13 are shown as we used 5-point image derivative. For feature-based methods, we were required to skip three frames for every estimate (so comparison ends at frame 12) to make the estimation result more stable. The averaged angular errors of the proposed method **AFS+AFM** are 2.741° for translation and 1.850° for rotation, with magnitude error 14.83%. We achieved similar motion accuracy as **TV-L_1-NL+LQP**. Our average runtime is 5.989 *sec/frame*. This is 4.375 times

faster than **spher-5-pt RANSAC** and 59.81 times faster than **TV-L_1-NL+LQP**.

7. Conclusion

We have proposed two constraints that allow normal flows to be used directly for motion recovery, which are readily usable with the availability of wide field-of-view imaging system. The first constraint separates the directions of translation and rotation components from general motion. The second constraint refines the solution set further and recovers the magnitude of rotation.

The proposed direct method has the advantage that it has tight constraint on the motion parameters. It requires neither distinctly trackable features in the video, piecewise smooth scene, nor interpolating the flow field. Synthetic experiment shows that the proposed method is robust against noise. Real image experiment reveals that the proposed method has similar motion accuracy as the state-of-the-art methods but having lower computational complexity.

References

[1] Y. Aloimonos and Z. Duric. Estimating the heading direction using normal flow. *IJCV*, 13(1):33–56, 1994.

[2] P. Baker, R. Pless, C. Fermüller, and Y. Aloimonos. A spherical eye from multiple cameras (makes better models of the world). *CVPR*, pages 576–583, 2001.

[3] J.-V. Bouguet. http://vision.caltech.edu/bouguetj/calib_doc, 2011.

[4] T. Brodský, C. Fermüller, and Y. Aloimonos. Directions of motion fields are hardly ever ambiguous. *IJCV*, 26:5–24, 1998.

[5] T. Brodský, C. Fermüller, and Y. Aloimonos. Structure from motion: beyond the epipolar constaint. *IJCV*, 37(3):231–258, 2000.

[6] A. Bruhn and J. Weickert. Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. *IJCV*, 61(3):211–231, 2005.

[7] C. Fermüller and Y. Aloimonos. Direct perception of three-dimensional motion through patterns of visual motion. *Science*, 15:1973–1976, 1995.

[8] C. Fermüller and Y. Aloimonos. Qualitative egomotion. *IJCV*, 15:7–29, 1995.

[9] C. Fermüller and Y. Aloimonos. On the geometry of visual correspondence. *IJCV*, 23(3):223–247, 1997.

[10] C. Fermüller and Y. Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *IJCV*, 28(2):137–154, 1998.

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

[12] R. H. Hardin, N. J. A. Sloane, and W. D. Smith. Tables of spherical codes with icosahedral symmetry. http://research.att.com/∼njas/icosahedral.codes, 2011.

[13] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

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

[15] B. Horn and B. Shunck. Determining optical flow. *AI*, 17:175–203, 1981.

[16] B. Horn and E. Weldon. Direct methods for recovering motion. *IJCV*, 2:51–76, 1988.

[17] C. Hu and L. Cheong. Linear quasi-parallax SfM using laterally-placed eyes. *IJCV*, 84:21–39, 2009.

[18] T.-W. Hui and R. Chung. Determining spatial motion directly from normal flow: A comprehensive treatment. *ACCV 2010 Workshops*, pages 23–32, 2011.

[19] J. Kim, M. Hwangbo, and T. Kanade. Spherical approximation for multiple cameras in motion estimation:Its applicability and advantages. *CVIU*, 114(10):2068–2083, 2010.

[20] J. Kim, H. Li, and R. I. Hartley. Motion estimation for nonoverlapping multicamera rigs: Linear algebraic and L∞ geometric solutions. *TPAMI*, 32(6):1044–1059, 2010.

[21] L. Kneip, R. Siegwart, and M. Pollefeys. Finding the exact rotation between two images independently of the translation. *ECCV*, pages 696–709, 2012.

[22] P. D. Kovesi. http://csse.uwa.edu.au/∼pk/research/matlabfns/, 2012.

[23] H. Li and R. Hartley. Five-point motion estimation made easy. *ICPR*, pages 630–633, 2006.

[24] J. Lim and N. Barnes. Directions of egomotion from antipodal points. *CVPR*, pages 1–8, 2008.

[25] J. Lim and N. Barnes. Estimation of the epipole using optical flow at antipodal points. *CVIU*, 114:245–253, 2010.

[26] M. I. A. Lourakis. Egomotion estimation using quadruples of collinear image points. *ECCV*, pages 834–848, 2000.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[28] A. Makadia, C. Geyer, and K. Daniilidis. Correspondence-free structure from motion. *IJCV*, 73(3):311–327, 2007.

[29] S. Mehta and R. Etienne-Cummings. A simplified normal optical flow measurement CMOS camera. *Trans. on Circuits and Systems I*, 53(6):1223–1234, 2006.

[30] R. C. Nelson and J. Aloimonos. Finding motion parameters from spherical motion fields (or the advantages of having eyes in the back of your head). *Biological Cybernetics*, 58:261–273, 1988.

[31] D. Nister. An efficient solution to the five-point relative pose problem. *TPAMI*, 26(6):756–770, 2004.

[32] R. Pless. Using many cameras as one. *CVPR*, pages 578–593, 2003.

[33] F. Raudies and H. Neumann. An efficient linear method for the estimation of ego-motion from optical flow. *DAGM Sym. on PR*, pages 11–20, 2009.

[34] C. Silva and J. Santos-Victor. Robust egomotion estiamtion from the normal flow using search subspaces. *TPAMI*, 19(9):1026–1034, 1997.

[35] M. V. Srinivasan and S. W. Zhang. Visual motor computations in insects. *Annual review of neuroscience*, 27:679–696, 2004.

[36] D. Sun, S. Roth, and M. J. Black. Secrets of optical flows estimation and their principles. *CVPR*, pages 2432–2439, 2010.