Pattern Recognition Letters 52 (2015) 59–64

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Camera motion estimation through monocular normal flow vectors ☆

Ding Yuan [a], Miao Liu [a], Jihao Yin [a,*], Jiankun Hu [b]

[a] School of Astronautics, Beihang University, Beijing 100191, China
[b] School of Engineering and Information Technology, University College, The University of New South Wales, Australian Defense Force Academy, Canberra ACT 2600, Australia

ARTICLE INFO

Article history:
Received 11 February 2014
Available online 13 October 2014

Keywords:
Camera motion estimation
Normal flow
Monocular
Optical flow

ABSTRACT

In this paper, we propose a method to directly estimate a camera's motion parameters by using normal flow vectors. In contrast to traditional methods, which tackle the problem by calculating optical flows or establishing motion correspondences, our proposed approach does not require conventional assumptions about the captured scene, such as consistent smoothness or distinct feature availability. In the proposed algorithm, the normal flows are classified into different groups, and each group will provide a possible solution regarding the camera's motion parameters. Then, the strategy of hypothesis and confirmation is adopted to eliminate the incorrect solutions. Finally, the optimal solution is obtained via the clustering algorithm. We have tested the proposed method on both synthetic image data and real image sequences. The experimental results illustrate the feasibility and reliability of the algorithm.

1. Introduction

Camera motion estimation, which estimates the relative motion between an observer and an object, is an important research topic in the areas of computer vision and image coding. It is an important feature in video sequences and is often a preliminary step for many computer vision algorithms and video data processing techniques such as vehicle navigation, human–computer interaction, video data structuring, video stabilization, video retrieval, and video abstraction. However, it is still a very challenging task despite decades of research because of its ill-posed characteristic. The motion is described in 3D, while the image is the projection of a 3D scene onto a 2D plane.

There have been abundant published approaches to camera motion estimation, which can be approximately classified into three categories. One can be summarized as the correspondence-based method, in which distinct features are extracted and tracked to establish their correspondences across consecutive frames. The matched features can describe the motion correspondence and can be used to estimate the epipolar geometry and the fundamental matrix. Finally, the camera motion parameters can be obtained by decomposing the fundamental matrix [2,5]. However, establishing accurate feature correspondences is itself a very challenging task. Another major category is the gradient category, in which the camera's ego-motion is frequently derived from the full optical flows [13,15]. The early works can be found in [4,14]. However, calculating the optical flow from an image sequence is another famous ill-posed problem in the field of computer vision because of the well-known aperture problem. Generally, we have to turn to some artificial constraints, such as it is smooth everywhere or the image domain is continuous and differentiable in space and time, to infer the optical flows. These artificial constraints do not often apply to real scenes.

In addition to the two major categories of approaches, there is another category of methods that tackle the camera motion problem by directly using normal flows. The normal flow is the projection of the optical flow along the direction of the intensity gradient. It describes the spatio-temporal gradient variance of the image sequence and can be calculated directly from the image sequence without any artificial constraints.

Aloimonos and Duric proposed a method on estimating the camera heading direction by using normal flow vectors [1]. However, only the camera's translation was discussed on the condition that the rotation of the camera is known beforehand. In Refs. [8,9], Fermüller and Aloimonos proposed a motion vector field and assigned each eligible pixel with its normal flow a positive or negative label. The boundaries that separate the positive labels and negative labels within the image domain implied the motion parameters of the camera. The motion vector field has been described in a planar model [8] and a spherical model [9], respectively. However, locating the boundaries precisely on the pattern consisting of sparse positive labels and negative labels is a very challenging task. Silva and Santos-Victor [16] achieved the motion estimation by searching subspaces and finding the lines con-

☆ This paper has been recommended for acceptance by D. Coeurjolly.

* Corresponding author. Tel.: +86 10 8231 6502; fax: +86 10 8233 8798.
E-mail addresses: dyuan@buaa.edu.cn (D. Yuan), mliu@sa.buaa.edu.cn (M. Liu), yjh@buaa.edu.cn (J. Yin), j.hu@adfa.edu.au (J. Hu).

http://dx.doi.org/10.1016/j.patrec.2014.09.015
0167-8655/

taining the Focus of Expansion (FOE). The intersection of two different lines can be viewed as the FOE, which represents the direction of the translation. The rotational parameters were inferred from the coefficients of the lines. However, any tiny tolerance in calculating the slope of the line would lead to a large error in estimating the motion parameters, as the method excessively depends on the magnitude of the normal flow, which is very sensitive to noise compared to the direction of the normal flow. In Ref. [6], Drareni and Martin proposed a method with an assumption that the depth information can be represented as an arbitrarily constant, on the condition that the depth of the captured scene is of uniform distribution. Although the motion parameters could be directly calculated from the image intensities, the assumption about the depth information restricts the method to very limited applications. Inspired by the structure of insects' vision systems, Hui and Chung determined the motion parameters directly from normal flows on the spherical eye platform [11]. They simulated the spherical eye platform by using a multi-camera rig, and the motion parameters could be obtained by analyzing the normal flow vector fields observed from multiple viewpoints. However, the method does not work with the more popular monocular vision system.

In contrast to the approaches mentioned above, in this paper we propose a novel method to estimate camera motion parameters by directly using normal flows under a monocular vision system. We group the normal flows into different categories according to their geometric characteristics and calculate the possible motion parameters for each group of normal flows. We suppose the camera undergoes the motion within the possible motion parameters and inspect whether the possible motion parameters are adapted to the normal flow vector field throughout the image domain. The optimal solution will be obtained by adopting our strategy of hypothesis and confirmation. Our method has the same advantages as the other approaches that tackle the problem by using normal flows; we neither require the captured scene to be smooth or differentiable nor demand a large amount of distinct features appear in the image. Moreover, our algorithm is independent of the texture information of the image frame. It is able to handle cases where the image domain is not highly textured, though the highly textured frame would help to improve the computational accuracy.

In this research, only the estimation of the relative motion between a moving observer and a static scene is of interest. The paper is organized as follows. Section 2 introduces the background briefly and explains the definitions of some particular normal flows. Section 3 presents our novel algorithm for camera motion estimation. The experimental results including synthetic image data and real image data are shown in Section 4. Section 5 concludes the proposed work and recommends future work.

2. Background and definitions

In this work, we assume a static scene and a monocular moving camera. Suppose that O is the optical center of the camera, and the Z-axis, perpendicular to the image frame, is defined as the optical axis. We suppose the camera is translating with the translational vector $\mathbf{t} = (U\ V\ W)^{\mathrm{T}}$ and rotating about the rotational axis $\boldsymbol{\omega} = (\omega_1\ \omega_2\ \omega_3)^{\mathrm{T}}$ that passes through the optical center. For a 3D scene point $\mathbf{P} = (X\ Y\ Z)^{\mathrm{T}}$ with respect to the camera coordinate ($OXYZ$), its projection on the image frame is $\mathbf{p} = (x\ y)^{\mathrm{T}}$ under the perspective camera model. Then, the optical flow $\mathbf{u} = [u,\ v]^{\mathrm{T}}$ at pixel $\mathbf{p}$ can be calculated with the following famous equation:

$$\begin{cases} u = \dfrac{(-fU + xW)}{Z} + \omega_1 \dfrac{xy}{f} - \omega_2 \left(\dfrac{x^2}{f} + f\right) + \omega_3 y \\ v = \dfrac{(-fV + yW)}{Z} + \omega_1 \left(\dfrac{y^2}{f} + f\right) - \omega_2 \dfrac{xy}{f} - \omega_3 x \end{cases} \tag{1}$$

where f is the focal length of the camera.

The optical flow can be considered as a summation of two components, the translational component and the rotational component, which exactly correspond to the camera's translation and rotation, respectively. Then, Eq. (1) can be rewritten as (i) in the vector form for a simplified expression:

$$\begin{aligned} \mathbf{u}(\mathbf{p}) &= \mathbf{u}_{\mathrm{trans}}(\mathbf{p}) + \mathbf{u}_{\mathrm{rot}}(\mathbf{p}) \\ &= \frac{W}{Z(\mathbf{p})}(\mathbf{p} - \mathrm{FOE}) + \mathbf{R}(\mathbf{p})\boldsymbol{\omega} \end{aligned} \tag{i}$$

where $\mathbf{u}(\mathbf{p})$ is the optical flow at pixel $\mathbf{p}$, $\mathbf{u}_{\mathrm{trans}}(\mathbf{p})$ and $\mathbf{u}_{\mathrm{rot}}(\mathbf{p})$ denote the translational component and the rotational component, respectively, and $Z(\mathbf{p})$ represents the depth information. The Focus of Expansion (FOE), the intersection of the translational vector $\mathbf{t}$ and the image plane if the camera undergoes a forward translation, is expressed as $FOE = (fU/W\ \ fV/W)^T$. The camera motion t can also be described as an intersection point of all translation trajectories, the Focus of Contraction (FOC), if the camera makes a backward translation. The coefficient matrix $\mathbf{R}(\mathbf{p})$ of the rotation parameter $\boldsymbol{\omega}$ is:

$$\mathbf{R}(\mathbf{p}) = \begin{bmatrix} xy/f & -(x^2/f + f) & y \\ (y^2/f + f) & -xy/f & -x \end{bmatrix} \tag{2}$$

Inferring the camera motion parameters from the optical flows calculated from the image sequence seems like a straightforward solution. However, similar to the matching problem in the field of computer vision, the problem of estimating optical flows is another ill-posed problem. The unique solution can only be achieved by adopting some artificial constraints, such as the smoothness constraint or the discontinuity constraint. Moreover, the accuracy of the calculation of optical flows depends on whether the assumed artificial constraints are applicable to the captured scene. Therefore, the approaches that estimate the camera's motion parameters based on the calculation of the optical flows often fail if the scene is not smooth everywhere, or the structure of the scene is not continuous and differentiable in space and time.

Normal flows, the projections of the optical flows along the direction of the gradient of image intensity, could be calculated directly from the image sequence without any artificial assumptions about the captured scene. In this work, we assume that the scene is stationary and the camera undergoes an arbitrary motion including translation and rotation. Then, the normal flows calculated from the image sequence contain the information of camera's translation, camera's rotation, depth of the scene and the gradient of the image intensity. It seems that determining the camera's motion directly from the normal flows is an unsolvable problem, as all of the data mentioned above are mixed together. However, there are some normal flows pointing in some particular directions along which the information of camera's translation can be exactly eliminated. Similarly, there are also some normal flows that do not carry the information about the camera's rotation. Then, the problem of estimating camera motion can be converted to the question on how to select those specific normal flows and how to utilize the normal flows to obtain the camera's motion parameters.

Assume that $\hat{\mathbf{n}}(\mathbf{p})$ is the normalized gradient of the image intensity at pixel $\mathbf{p}$. Then, the magnitude of the normal flow $V^n(\mathbf{p})$ at $\mathbf{p}$ is defined as:

$$V^n(\mathbf{p}) = \mathbf{u}(\mathbf{p}) \cdot \hat{\mathbf{n}}(\mathbf{p}) = V^n_{\mathrm{trans}}(\mathbf{p}) + V^n_{\mathrm{rot}}(\mathbf{p}) \tag{3}$$

where $V^n_{\mathrm{trans}}(\mathbf{p}) = \frac{W}{Z(\mathbf{p})}\hat{\mathbf{n}}(\mathbf{p})(\mathbf{p} - \mathrm{FOE})$, and $V^n_{\mathrm{rot}}(\mathbf{p}) = \hat{\mathbf{n}}^{\mathrm{T}}(\mathbf{p})\mathbf{R}(\mathbf{p})\boldsymbol{\omega}$.

In Eq. (3), similar to the definition of the optical flow, the normal flow can be treated as a summation of two components: the translational component $V^n_{\mathrm{trans}}(\mathbf{p})$ and the rotational component $V^n_{\mathrm{rot}}(\mathbf{p})$. Calculating the motion parameters directly from Eq. (3) is an ill-posed problem, as the depth information $Z(\mathbf{p})$ is unknown, even if the pixels with their corresponding normal flows are given. However, if there exist some pixels whose normal flows are pointing at particular directions that can completely eliminate the translational components