Pattern Recognition 48 (2015) 422–437

Contents lists available at ScienceDirect

Pattern Recognition

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

Determining shape and motion from monocular camera: A direct approach using normal flows

Tak-Wai Hui [a,*], Ronald Chung [b]

[a] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong
[b] Vocational Training Council of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 26 November 2012
Received in revised form
24 May 2014
Accepted 14 August 2014
Available online 26 August 2014

Keywords:
Camera motion
Direct method
Normal flow
Optical flow

ABSTRACT

Determining the spatial motion of a moving camera from a video is a classical problem in computer vision. The difficulty of this problem is that the flow pattern directly observable in the video is generally not the complete flow pattern induced by the motion, but only the partial information of it, which is known as the normal flow. In this paper, we present a direct method which neither requires the establishment of feature correspondences nor the recovery of optical flow between two image frames, but we directly utilize all observable normal flow data to recover the camera motion. We propose a two-stage iterative algorithm to search the solution in the motion space in a coarse-to-fine framework. The first stage involves the use of the direction part of the normal flow. Each of these normal flow data can provide a constrained solution space to the direction of motion. The intersection of the motion solutions from all the available normal flow data can reduce the motion ambiguity to a certain extent. We then use the globality of the rotational magnitude to all image positions to constrain the motion parameters further. Once the camera motion is determined, the depth map of the imaged scene (up to an arbitrary scale) can be recovered. Experimental results on synthetic data and real images are provided to reveal the performance of the proposed method.

1. Introduction

A camera which moves in a static scene generally induces a certain apparent flow pattern in the acquired video. The capabilities of revealing the position, moving direction, and other dynamic information from the video are essential for higher level tasks such as autonomous navigation, visual control, human action understanding, and more. Due to the well-known ambiguity between motion speed and object size-and-depth, from monocular video alone, the translation magnitude of the motion is generally not determinable and left as an overall arbitrary scale related to object depth. The other motion parameters are determinable. In other words, if we describe the spatial motion as consisting of a translation component $\mathbf{t}$ (a 3-vector, whose direction and magnitude represent the direction and magnitude of the translation in space respectively) and a rotation component $\mathbf{w}$ (also a 3-vector, which represents the rotation in space in angle-axis form), our task is to determine the direction of $\mathbf{t}$ and the full $\mathbf{w}$.

The usual approaches to determine motion parameters are based either on establishing feature correspondences [1,2], optical flows (also known full flows) [3–7], or normal flows (so-called direct methods) [8–13], between two image frames in a video.

The correspondence-based methods require tracking of distinct features which might not be always available in the video. The existence of repetitive patterns can cause ambiguity in establishing the correct correspondences. On the other hand, the optical flow induced by the spatial motion at any image position is observable generally only partially. The apparent flow, termed the normal flow, is the component of the full flow along or opposite to the direction of the local intensity gradient. The partial observability of the flow is what makes motion determination a challenge. Recovery of optical flow often requires piecewise-smooth flow. However, this assumption is not valid near depth discontinuities. The state-of-the-art methods (such as [14–17]) recover optical flow by minimizing an energy functional. An accurate optical flow field cannot be obtained unless there is a good compromise between the data term and the regularization term. Interpolation of optical flow from textured image region to homogeneous image region is often necessary. However, the minimization process is computationally expensive and often requires the use of graphical processing unit (GPU) or multi-core central processing unit (CPU) in order to achieve real-time performance. For instance, a recent work used 7 min to compute the optical flow for an image pair having resolution 640×480 on a rather high performance laptop computer [17]. On the contrary,

* Corresponding author.
E-mail addresses: twhui1@mae.cuhk.edu.hk (T.-W. Hui), rchung@vtc.edu.hk (R. Chung).

http://dx.doi.org/10.1016/j.patcog.2014.08.012
0031-3203/

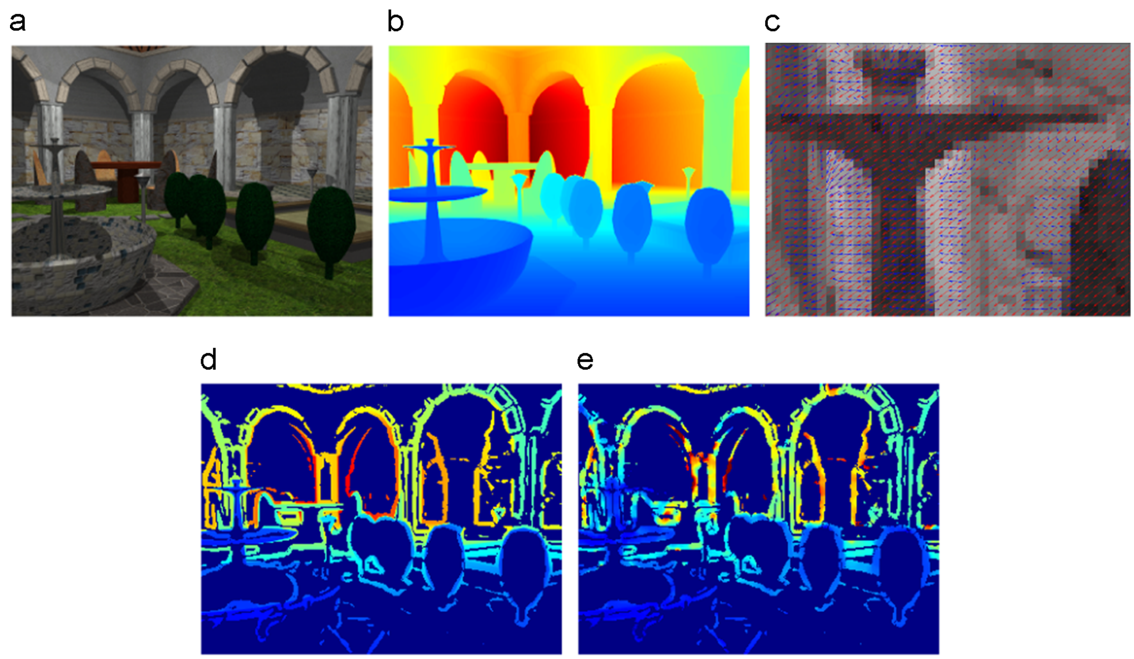

Fig. 1. (a) An image frame in the Fountain sequence [7]. (b) The ground-truth depth map. Closer parts appear cooler color in the image. (c) The ground-truth optical flow vectors (red arrows) and the actually computed normal flow vectors (blue arrows) near the top of the fountain. (d) The recovered depth map using the normal flow by projecting the ground-truth optical flow along local intensity gradient. (e) The recovered depth map using the actually computed normal flow. The depth map is denoised by a median filter (kernel size: 9×9). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the direct methods determine the motion parameters from normal flow which is directly measurable from the spatial-temporal intensity gradient [18]. They demand relatively less computation resource than those methods based on feature correspondence and optical flow. Moreover, normal flow also can be measured by custom-designed vision sensor directly [19]. Due to the above advantages, we investigate the use of normal flow to determine motion parameters.

In Fig. 1, we show the recovered depth maps of a frame in the Fountain sequence [7] using different normal flows with the known camera motion. We first project the ground-truth optical flow along the local intensity gradient to form the normal flow. Fig. 1c illustrates some parts of the optical flow field and the normal flow field. It should be noticed that no normal flow exists in some parts of the image space because the spatial-temporal intensity gradients are very weak there. Fig. 1d shows the resulted depth map which is very close to the ground truth. Fig. 1e illustrates the recovered depth map using the normal flow computed from the spatial-temporal intensity gradient without prior knowledge about the full flow. We can also observe that the errors in the depth map are relatively larger at the image positions corresponding to the regions of the scene that are far away from the camera. The only difference between the above two normal flows is the difference sources of the temporal intensity gradient. Since the direction part of normal flow mainly depends on the spatial intensity gradient,[1] we can conclude that most of the normal flow extraction errors are indeed originated from the magnitude part. This can be also explained from the fact that the spatial resolution of an image sequence is generally higher than its temporal resolution. This provides us the insight that postponing the use of the magnitude component of normal flow can improve the accuracy and robustness for determining the spatial motion.

In this article, we provide a direct method to determine the camera motion in a video. Some preliminary results have been published in our previous work [20]. Our contribution is three-fold. First, we separately utilize the direction and magnitude components of normal flow to determine the camera motion. This makes the motion estimation more robust to noise. Second, the separation of two components facilitates the development of the two constraints. One is related to the direction component of the normal flow – the Apparent Flow Direction (AFD) constraint, and the other to the globality of the rotational motion magnitude at all image positions – the Apparent Flow Magnitude (AFM) constraint. The AFD constraint manifests itself as a system of linear inequalities that bind the motion parameters using only the direction component of the flow field. The AFM constraint serves to reduce motion ambiguity further by insisting that every image position must has a component of normal flow magnitude that is consistent with the same rotation magnitude of the spatial motion. Third, we make the motion estimation process more computationally efficient by exploiting the two constraints in a two-stage iterative voting process using a coarse-to-fine framework.

2. Related works

2.1. Optical flow

Starting from the seminal works by Horn and Shunck [21] and Lucas and Kanade [22], many solutions have been proposed for dealing with the shortcomings of previous models. Global approach such as the work from Horn and Shunck [21] yields optical flow with full density, but it is experimentally known to be more sensitive to noise. Local method such as the work from Lucas and Kanade [22] is relatively more robust under noise, but it does not give dense flow field. Bruhn et al. combined the local and global methods to give a compromise between the two approaches [23]. Weickert and Schnörr [24] extended the spatial flow-driven

[1] Strictly speaking, the direction of each normal flow vector depends on the spatial intensity gradient and the sign of the temporal intensity gradient. But if we remove those normal flow vectors which have very small temporal gradient magnitude, then the direction parts are error-free from the temporal gradient. Therefore, it is safe to say that the direction part of normal flow mainly depends on the spatial gradient.