J. Vis. Commun. Image R. 25 (2014) 98–107

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Local scene flow by tracking in intensity and depth

Julian Quiroga [a,b,*], Frédéric Devernay [a], James Crowley [a]

[a] INRIA Grenoble Rhone-Alpes, 655 avenue de l'Europe, 38334 Saint Ismier Cedex, France
[b] Departamento de Electrónica, Pontificia Universidad Javeriana, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 17 July 2012
Available online 1 April 2013

Keywords:
Scene flow
3D motion estimation
Image tracking
Brightness consistency
Depth data
Locally-rigid motion
Optical flow
Image warping

ABSTRACT

The scene flow describes the motion of each 3D point between two time steps. With the arrival of new depth sensors, as the Microsoft Kinect, it is now possible to compute scene flow with a single camera, with promising repercussion in a wide range of computer vision scenarios. We propose a novel method to compute a local scene flow by tracking in a Lucas–Kanade framework. Scene flow is estimated using a pair of aligned intensity and depth images but rather than computing a dense scene flow as in most previous methods, we get a set of 3D motion vectors by tracking surface patches. Assuming a 3D local rigidity of the scene, we propose a rigid translation flow model that allows solving directly for the scene flow by constraining the 3D motion field both in intensity and depth data. In our experimentation we achieve very encouraging results. Since this approach solves simultaneously for the 2D tracking and for the scene flow, it can be used for motion analysis in existing 2D tracking based methods or to define scene flow descriptors.

1. Introduction

The scene flow corresponds to the 3D motion field of the scene [1] and since it provides the motion of 3D points, an accurate estimation of the scene flow can be useful in a wide variety of applications including navigation, interaction, object segmentation, motion analysis, tracking, *etc*.

A current topic of great interest is human activity understanding, where video analyzing and interpretation is required to perform recognition or classification, and the 3D information given by the scene flow may be used to provide powerful features. However, there is no work that directly computes scene flow to perform tasks like human action recognition or gesture classification. Probably, this is due to the fact that most existing methods require stereo or multi-view camera systems, which are not always available. Besides, most of these methods compute a dense scene flow by optimizing a global energy function, spending a lot of processing time and becoming not suitable for real time applications. On the other hand, the optical flow that is related with the scene flow projection on the image plane has been successfully used in human action recognition. Histograms of optical flow are commonly used in state-of-the-art techniques in action recognition to construct descriptors over spatio-temporal interest points [2–4] and to extract 2D trajectories by tracking key-points [4–6]. Furthermore, since trajectory based methods outperform other state-of-the-art approaches for action classification [7], it is promising to use scene flow to capture motion information by extracting accurate 3D trajectories.

Recently with the arrival of depth cameras, based either on time-of-flight (ToF) or structured light sensing, it has been possible to compute scene flow using a pair of registered sequences of depth and intensity, as recently in [8,9]. Depth sensors have decreased system requirements needed to compute the 3D motion field, opening the door to incorporate scene flow based features in common recognition tasks. Some attempts have been made to include depth data in human action recognition tasks. For example, a bag of 3D points extracted from the depth data is used for recognition in [10] while in [11] descriptors obtained by well known techniques [2,12] were extended with depth information, outperforming the original methods. Similarly, when a depth sensor is available the scene flow can be inferred from the optical flow by using the depth information. However, as we show in this paper, even small errors in the optical flow may generate significant errors in the scene flow computation. Computing scene flow in this way does not fully exploit the relation between the intensity and depth information. In this work we aim to explore how to simultaneously use intensity and depth data to compute local scene flow. As a result, we propose a method that can be used to get accurate 3D trajectories and define scene flow based descriptors.

One of the main contributions of this paper is the definition of a pixel motion model that allows the constraint of the scene flow in the image. Using this motion model and assuming a 3D local rigidity of the scene, we are able to solve for the scene flow that best

[*] Corresponding author at: INRIA Grenoble Rhone-Alpes, 655 avenue de l'Europe, 38334 Saint Ismier Cedex, France.

E-mail addresses: julian.quiroga@inria.fr (J. Quiroga), frederic.devernay@inria.fr (F. Devernay), james.crowley@inria.fr (J. Crowley).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.03.018

explains the observed intensity and depth data. Therefore, our method combines information of both sensors and simultaneously solve for the scene flow and its projection in the image, which we named *image flow*. This approach differs from previous scene flow methods using depth sensors, since they reconstruct the scene flow from the observed optical flow [8] or by using a large number of hypotheses to explain the motion of each point in 3D [9], without exploiting the 2D parameterization. Moreover, unlike other scene flow methods that suffer from the smoothness constraint brought by 2D parameterization, we use a 3D local rigidity assumption, which is approximately real for most of the scenes of interest.

The other contribution of the paper is the formulation of a local scene flow computation method by extending the Lucas–Kanade framework [13] to exploit both intensity and depth data. In this formulation, it is possible to treat with large displacements in a coarse-to-fine procedure. Besides, instead of solving for a dense scene flow by optimizing a global energy, as most of previous methods, we solve for a local scene flow that can be focused over a selected set of key-points. This formulation is versatile enough to extract accurate 3D trajectories, initialize other methods by computing a dense scene flow, or refine a estimated 3D motion field over specific points. Unlike previous scene flow methods our local approach is suitable for real time applications and its extension to multiple cameras or depth sensors is straightforward.

1.1. Related work

Scene flow was first introduced by Vedula et al. [1] as the full 3D motion field in the scene. Most scene flow methods assume a stereo, or multi-view camera system, in which the motion and the geometry of the scene are jointly estimated, in some cases, under a known scene structure. Since optical flow is (an approximation of) the projection of the 3D motion field on the camera image plane, an intuitive way to compute scene flow is to reconstruct it from the optical flow measured in a multi-view camera system, as proposed by Vedula et al. [14], or including a simultaneously structure estimation as Zhang and Kambhamettu [15]. However, it is difficult to recover a scene flow compatible with several observed optical flows that may be contradictory.

The most common approach for estimating scene flow is to perform an optimization on a global energy function, including photometric constraints and some regularization. Some authors introduce constraints of a full calibrated stereo structure [16–20]. Wedel et al. [17] enforce consistency on the stereo and motion solution but they decouple the disparity at the first time step without exploiting the spatio-temporal information. To overcome this limitation, simultaneous solution of the scene flow and structure was proposed. Huguet and Devernay [18] simultaneously compute the optical flow field and two disparities maps, while Valgaerts et al. [21] assume that only the camera intrinsics are known and they show that scene flow and the stereo structure can be simultaneously estimated.

All these methods suffer from the smoothness constraints brought by 2D parametrization. Basha et al. [19] improve the estimation by formulating the problem as a point cloud in 3D space and the scene flow is regularized using total variation (TV). Recently, Vogel et al. [20] regularize the problem by encouraging a locally rigid 3D motion field, outperforming TV regularization. Furthermore, other methods simultaneously solve the 3D surface and motion [22,23]. Another possibility is to work in the scene domain, and to track 3D points or surface elements [24,25]. Carceroni and Kutulakos [24] model the scene as a set of surfels but it requires a well-controlled lighting and acquisition setup, and because its complexity the scene flow solution is only suitable in a limited volume. Rather than computing a dense scene flow, Devernay et al. [24] directly get a set of 3D trajectories from which the scene flow is derived. However, this method suffers from drifts problems and its proposed point visibility handling is a difficult task.

When a depth camera is available, the sensor provides structure information and surface estimation is not needed. Spies et al. [26] estimate the scene flow by solving for the optical flow and range flow. Lukins and Fisher [27] extend this approach to multiple color channels and one aligned depth image. In these approaches the 3D motion field is computed by constraining the flow in intensity and depth images of an orthographically captured surface, so that, the range flow is not used to support the optical flow computation. Letouzey et al. [8] directly estimate the 3D motion field using photometric constraints and a global regularization term without fully exploiting the information given by the depth sensor. Recently, Hadfield and Bowden [9] estimate the scene flow by modeling moving points in 3D using a particle filter, reducing the oversmoothing caused by global regularization. However, this method requires a lot of computational time since a large number of motion hypotheses must be generated and tested for each 3D point.

1.2. Our approach

Similar to [8,9], we estimate the scene flow using a pair of aligned intensity and depth sequences. Although, rather than computing a dense scene flow we get a set of 3D motions by tracking in the image domain using a coarse-to-fine procedure. The work in this paper is inspired by that of Devernay et al. [25], in which a sparse scene flow is derived from 3D trajectories using several cameras. In our approach, instead of tracking 3D points we use a Lucas–Kanade framework [13] to solve for a local scene flow using constraints in intensity and depth data.

Previous works [26,27] solve at the same time for the optical flow and range flow assuming an orthographic camera, however, under this approach the estimated depth velocity can not be included to constraint the optical flow computation. Instead, we directly compute a local scene flow by tracking small surface patches in intensity and depth data. As in [20], we assume that the scene is composed of independently, but rigidly, moving 3D parts avoiding the use of smoothness constraints in 2D. Thus, considering a 3D local rigidity of the scene, we model the image flow induced by the surface motion by using a *rigid translation flow model*. This model allows the constraint of the 3D motion field in the image domain. In this way we are able to solve for the scene flow that best explains the observed intensity and depth data for each interest region on the image.

Previous Lucas–Kanade methods use a 2D warping [13]. Unlike them, we model the image flow as a function of the 3D motion vector with help from a depth sensor, improving the accuracy of the optical flow and solving directly for the scene flow. Besides, without expecting a planar surface patch as in surfels based techniques [24,25], this motion model allows the constraint of scene flow in intensity and depth images. In order to treat with large displacement the scene flow can be propagated in a coarse-to-fine strategy.

Incorporating depth data our approach improves tracking precision in the image domain that allows at the same time the computation of the scene flow. Because we solve directly for the scene flow by performing tracking in the image domain, this method can be directly used to extract accurate 2D or 3D trajectories, initialize/refine other scene flow methods or define scene flow based descriptors for motion analysis.

1.3. Paper structure

The remainder of this paper is organized as follows. We begin in Section 2 with the definition of a motion model that allows the constraint of the 3D motion vector in the image domain. In