Computer Vision and Image Understanding 150 (2016) 66–80

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Multi-reference combinatorial strategy towards longer long-term dense motion estimation

Pierre-Henri Conze[a,*], Philippe Robert[b], Tomás Crivelli[b], Luce Morin[c]

[a] ICube UMR 7357, Université de Strasbourg, CNRS, 300 boulevard Sébastien Brant, CS 10413, 67412 Illkirch Cedex, France
[b] Technicolor, 975 avenue des Champs Blancs CS 17616, 35576 Cesson-Sévigné, France
[c] INSA Rennes, IETR, UMR 6164, UEB, 20 avenue des Buttes de Coesmes, 35708 Rennes Cedex 7, France

ARTICLE INFO

Article history:
Received 7 August 2015
Revised 21 April 2016
Accepted 22 April 2016
Available online 20 May 2016

Keywords:
Long-term motion estimation
Dense matching
Multi-reference frames
Combinatorial integration
Motion trajectories
Video editing

ABSTRACT

This paper addresses the estimation of accurate long-term dense motion fields from videos of complex scenes. With computer vision applications such as video editing in mind, we exploit optical flows estimated with various inter-frame distances and combine them through multi-step integration and statistical selection (MISS). In this context, managing numerous combinations of multi-step optical flows requires a complexity reduction scheme to overcome computational and memory issues. Our contribution is two-fold. First, we provide an exhaustive analysis of available single-reference complexity reduction strategies. Second, we propose a simple and efficient alternative related to multi-reference frames multi-step integration and statistical selection (MR-MISS). Our method automatically inserts intermediate reference frames once matching failures are detected to re-generate the motion estimation process and re-correlates the resulting dense trajectories. By this way, it reaches longer accurate displacement fields while efficiently reducing the complexity. Experiments on challenging sequences reveal improved results compared to state-of-the-art methods including existing MISS schemes both in terms of complexity reduction and accuracy improvement.

1. Introduction

Estimating accurate long-term dense correspondence fields is a fundamental task for many computer vision applications. A key tool in this context is optical flow whose early formulations come from the early 80s (Horn and Schunck, 1981; Lucas and Kanade, 1981). Significant progress has been made to improve both robustness and spatial consistency of the flow by introducing respectively more sophisticated data models than the classical brightness constancy assumption and robust discontinuity-preserving smoothness constraints.

However, most of state-of-the-art optical flow estimators focus on estimating dense motion between two consecutive frames only. They seldom consider that sequences comprise series of images that are inter-related. When tackling motion estimation over a video sequence, object-based (Pérez et al., 2002) or sparse (Shi and Tomasi, 1994) motion estimation is usually sufficient (visual servoing, surveillance, gestural human-machine interface, video indexing...). However, other applications explicitly require a dense and long-term description of how the video content evolves in time. Such applications include scene segmentation (Brox and Malik, 2010; Lezama et al., 2011), trajectory analysis (Wang et al., 2011) or video editing tasks like 2D-to-3D video conversion (Cao et al., 2011) or graphic elements insertion where each pixel of a given area needs to be tracked over many frames to be properly replaced by the corresponding pixel of the inserted element. Thus, we focus on this challenging issue: how to construct dense fields of point correspondences over extended time periods?

Establishing dense long-term correspondences requires to compute dense motion fields between distant frames and therefore to simultaneously handle small and large displacements. Optical flow is the appropriate tool for this task but classical optical flow assumptions which may fail between consecutive frames are even less valid between non-consecutive frames. When dealing with multiple frames and their associated point correspondences, another key aspect is the temporal consistency of the flow vectors which must depict temporal smoothness along trajectories. In this context, several recent studies have extended optical flow to the purpose of (semi-)dense long-term motion estimation. State-of-the-art deals with consecutive optical flow concatenation (Brox and Malik, 2010; Sundaram et al., 2010; Wang et al., 2013), trajectorial regularization (Salgado and Sánchez, 2007; Werlberger et al., 2009), particle representation (Sand and Teller, 2008), subspace

* Corresponding author.
E-mail address: conze@unistra.fr (P.-H. Conze).

http://dx.doi.org/10.1016/j.cviu.2016.04.013

constraints (Garg et al., 2013; Irani, 2002) as well as multi-step strategies (Conze et al., 2013; 2014; Crivelli et al., 2012a; 2012b).

Optical flows estimated between consecutive frames can straightforwardly be concatenated to construct motion trajectories along a video sequence through Euler or Runge–Kutta temporal integration (Butcher, 2008). This strategy has been exploited in many works (Brox and Malik, 2010; Sundaram et al., 2010; Wang et al., 2013) but may lead to large error accumulation resulting in a substantial drift over extended periods of time. Results in the literature are generally reported on fairly short sequences and reliable tracks usually do not exceed thirty frames.

To limit motion drift, optical flow estimation has been extended from two frames to multiple frames via hard (Werlberger et al., 2009) or soft (Salgado and Sánchez, 2007) spatio-temporal constraints which penalize motion variations along trajectories. Despite these contributions, more sophisticated motion models have been investigated to deal with complex motion. Sand and Teller (2008) represent video motion using a set of particles that move across the sequence. To reach variable-length point trajectories, particles are sequentially propagated using optical flows computed between consecutive frames. Using such representation forsakes rigidity assumptions and motion model considerations which may fail in complex situations but does not achieve full density.

Since trajectories of points belonging to an object are correlated even with strong deformations, subspace constraints-based methods assume that the set of all flow fields reside in a low-dimensional subspace (Irani, 2002). Therefore, a low-rank space is built to constrain optical flow estimation which provides additional information to solve the ambiguity in regions that suffer from the aperture problem. Garg et al. (2013) perform dense multi-frame optical flow estimation in a variational framework using 2D trajectory subspace constraints (Garg et al., 2013). This approach generates dense trajectories starting from a reference frame in a non-rigid context. Trajectories are estimated close to a low-dimensional trajectory subspace built through Principal Component Analysis (PCA) or Discrete Cosine Transforms (DCT). Nevertheless, this method requires strong a priori assumptions on the scene content. Moreover, only trajectories starting from a fixed reference frame are considered. The computation of motion fields starting from subsequent frames and going back to the reference frame is not under consideration.

The alternative concept of multi-step flow (MSF), Crivelli et al. (2012a); 2012b) focuses on how to construct long-term dense fields of correspondences using multi-step optical flows, i.e., optical flows computed between consecutive frames or with larger inter-frame distances. MSF sequentially merges a set of displacement fields at each intermediate frame, up to the target frame. This set is obtained via concatenation of multi-step optical flows with displacement vectors already computed for neighbouring frames. Multi-step estimations can handle temporary occlusions since they can jump occluding objects. Contrary to Garg et al. (2013), MSF considers both trajectory estimation between a reference frame and all the images of the sequence (from-the-reference) and motion estimation to match each image to the reference frame (to-the-reference). Two set-ups can be then considered: information pushing from the reference frame or information propagation over each frame by pulling it from the reference frame.

Despite its ability to handle both scenarios, MSF has two main drawbacks. First, it performs the selection of displacement fields by relying only on classical optical flow assumptions such as the brightness constancy constraint that may fail between distant frames. Second, the candidate displacement fields are based on previous estimations. It ensures a certain temporal consistency but can also propagate estimation errors along the subsequent frames of the sequence, until a new available step gives a chance to match with a correct location again.

These limitations can be solved by considering the multi-step integration and statistical selection (MISS) introduced in Conze et al. (2013); 2014) for the estimation of from-the-reference and to-the-reference long-term dense motion correspondences between a reference frame I_{ref} and all the other frames I_n of a video sequence. Based on pre-computed multi-step optical flows, similarly to MSF (Crivelli et al., 2012a), MISS algorithm processes each pair of frames $\{I_{ref}, I_n\}$ via both multi-step integration and statistical selection. Multi-step integration builds a large set of candidate displacement fields via the generation of multiple motion paths made of concatenated multi-step optical flows. Then, the statistical selection consists in selecting among the resulting set of candidate displacement fields the optimal one based on statistics and spatial regularization.

The statistical selection performs the displacement field selection by studying the redundancy on the large candidate set resulting from multi-step integration. For distant frames, it provides a more robust indication than classical optical flows assumptions involved in MSF (Crivelli et al., 2012a). Moreover, contrary to MSF (Crivelli et al., 2012a) which sequentially relies on previously established correspondences, MISS algorithms independently process each pair of frames $\{I_{ref}, I_n\}$ to prevent error propagation. Temporal consistency is handled a-posteriori through robust temporal smoothness constraints (Conze et al., 2014).

Each time the multi-step integration stage processes a given pair $\{I_{ref}, I_n\}$, only a subset of all the possible motion paths between I_{ref} and I_n can be generated and kept in memory due to computational and memory issues. For instance, the number of possible motion paths for a distance of 30 frames and with steps 1, 2, 5 and 10 is... 5877241! Up to a few hundreds can be actually built and kept in memory with current computer capabilities. To avoid these issues, the multi-step integration stage must include a computational complexity reduction strategy to prevent a cumbersome exhaustive motion paths generation process. This complexity reduction scheme must cleverly select a subset of all possible motion paths to minimize the tracking failure probability while increasing the trajectory lifetime.

In this direction, we aim at covering and extending the spectrum of MISS introduced in Conze et al. (2013); 2014) in the context of long-term dense motion estimation. After a brief overview of the baseline method (Section 2), two main contributions are addressed. First, given the computational and memory issues mentioned above, we identify and study the available single-reference complexity reduction schemes adapted to the multi-step integration stage of MISS (Section 3). Second, we propose a new, simple and efficient complexity reduction strategy based on an automatic multi-reference frames processing (Section 4). It reaches longer accurate displacement fields while efficiently reducing the complexity. Its ability to go towards longer long-term dense motion estimation is assessed through comparisons with state-of-the-art methods on challenging sequences (Section 5).

2. Multi-step integration and statistical selection (MISS)

The baseline multi-step integration and statistical selection (MISS) method Conze et al. (2013); 2014) can be, at first glance, studied without any complexity reduction considerations. Let us overview both multi-step integration and statistical selection steps in the context of exhaustive motion path generation. For the sake of clarity, complexity reduction is addressed only starting from Section 3.

2.1. Multi-step integration

The multi-step integration aims at producing a large set of displacement fields between a reference frame I_{ref} and a given