Multi-step flow fusion: towards accurate and dense correspondences in long video shots

Tomás Crivelli, Pierre-Henri Conze, Philippe Robert,
Matthieu Fradet, Patrick Pérez
firstname.lastname@technicolor.com

Technicolor

With high quality editing of video shots of arbitrary duration in mind, we focus on this problem: how to construct accurate dense fields of correspondences over extended time periods using series of elementary optical flows. Highly elaborated optical flow estimation algorithms are at hand, and they were applied before for dense tracking by simple accumulation, however with unavoidable position drift. On the other hand, direct long-term point matching is more robust to such deviations, but is very sensitive to ambiguous correspondences. Why not combining the benefits of both approaches? Following this idea, we develop a *multi-step flow fusion* method that optimally generates a dense long-term displacement field by first merging several candidate estimation paths and then filtering the tracks in the spatio-temporal domain. Our approach permits to handle small and large displacements with improved accuracy and is able to recover a trajectory from temporary occlusions.

Consider a sequence of RGB images $\{I_n\}_{n:0\dots N}$. Let $\boldsymbol{d}_{n,m} : \Omega \to \mathbb{R}^2$ be a *displacement field* defined on the continuous rectangular domain Ω, such that for every $\boldsymbol{x} \in \Omega$ it corresponds a *displacement vector* $\boldsymbol{d}_{n,m}(\boldsymbol{x}) \in \mathbb{R}^2$ for the ordered pair of images $\{I_n, I_m\}$. Given a *reference image*, say I_0, point tracking is compactly represented by $\boldsymbol{d}_{0,m}(\boldsymbol{x}) \forall m : 1\dots N$ (*from-the-reference* correspondences), *i.e.*, the set of displacement fields from I_0 to the subsequent frames I_m. Instead, for propagating (pulling) information present at a key reference frame to the rest of the sequence it is more natural to deal with $\boldsymbol{d}_{n,0}(\boldsymbol{x}) \forall n : 1\dots N$ (*to-the-reference* correspondences).

We address the problem of estimating from-the-reference as well as to-the-reference long-term displacement fields from elementary optical flow fields. Temporal integration of successive optical flow fields using classic tools such as Euler and Runge-Kutta schemes is possible (for instance in [3, 4]) but flow estimation errors are inevitably accumulated through this process. A solution would be to estimate the direct displacements between the reference frame and the other frames. However the longer the distance in time between two frames, the more ambiguous the matching process. So-called large displacement dense matching algorithms deal either with fast motions between consecutive frames [1] (but are not at all oriented to finding point correspondences along hundreds of frames) or assume parametric models [5] also constrained to limited frame distances. However, matching non-consecutive (time distant) frames can still be very useful as its accuracy much depends on inter-frame motion range: indeed one observes that for short/mid-term dense point matching, some regions of the image are better matched by concatenating consecutive motion vectors, while for others a direct matching is preferred (e.g., if displacement between consecutive frames is small). So, the idea is to consider multiple displacement fields with various inter-frame distances in order to have the best vectors available among all the candidates. The process is carried out in three phases: considering a pair $\{I_n, I_m\}$, first elementary optical flow fields with various inter-frame distances (called steps) are estimated. Then, various candidate displacement fields $\boldsymbol{d}_{n,m}$ are computed by different concatenations of the elementary fields, and finally the displacement field is obtained by merging these candidate fields. This is called Multi-Step Fusion (MSF).

Let us consider the pair $\{I_n, I_0\}$, corresponding to respectively the current and reference frames. Suppose that as an input we are given a set $\boldsymbol{M}_n$ of Q_n elementary optical flow fields $\boldsymbol{v}_{n,t}$ (between frames n and t): $\boldsymbol{M}_n = \{\boldsymbol{v}_{n,n+s_1}, \boldsymbol{v}_{n,n+s_2}, \dots, \boldsymbol{v}_{n,n+s_{Q_n}}\}$ for image I_n. Considering any step s_k, one can compute a set of displacement vectors between I_n and I_0 resulting from the combination of the elementary vector $\boldsymbol{v}_{n,n+s_k}$ and the displacement $\boldsymbol{d}_{n+s_k,0}$ available between I_{n+s_k} and I_0:

$$\boldsymbol{d}^k_{n,0}(\boldsymbol{x}) = \boldsymbol{v}_{n,n+s_k}(\boldsymbol{x}) + \boldsymbol{d}_{n+s_k,0}(\boldsymbol{x} + \boldsymbol{v}_{n,n+s_k}(\boldsymbol{x})). \tag{1}$$

In this manner we generate different candidate displacements or *paths* (Fig. 1) among which we aim at deciding the optimal for each pixel $\boldsymbol{x}$.

The selection of the optimal path for all the points of the grid for a pair $\{I_n, I_0\}$ is achieved via an appropriate global optimization stage that fuses all the candidate fields into a single optimal displacement field. To do so, we apply the method recently presented in [2] in the context of instantaneous optical flow estimation by flow fusion.

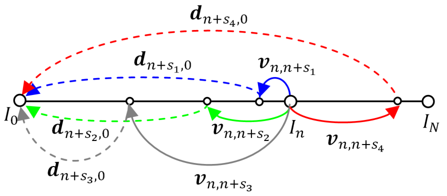
$\boldsymbol{d}_{n+s_4,0}$
$\boldsymbol{d}_{n+s_1,0}$
$\boldsymbol{v}_{n,n+s_1}$
I_0
$\boldsymbol{d}_{n+s_2,0}$
$\boldsymbol{v}_{n,n+s_2}$
I_n
$\boldsymbol{v}_{n,n+s_4}$
I_N
$\boldsymbol{d}_{n+s_3,0}$
$\boldsymbol{v}_{n,n+s_3}$

Figure 1: Multi-step point correspondence. The displacement from frame I_n to frame I_0 can be generated following different *paths* according to the available elementary motion fields (solid lines) and the previously estimated long-term displacements (dashed lines).

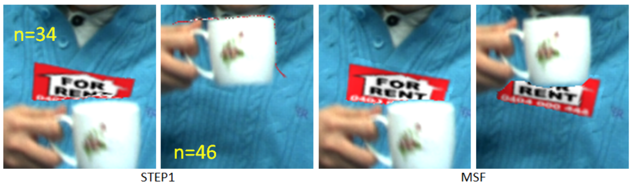

Figure 2: The *for rent* logo was inserted at frame I_0 by the user and was then automatically inserted in the other frames. The proposed MSF method overcomes the large occlusion by the arm while temporal integration of single step (equal to 1) optical flows (STEP1) fails.

At the end of the multi-step fusion stage, the set of forward vectors $\boldsymbol{d}_{0,n}(\boldsymbol{x})$ that give the position of point, originally at $\boldsymbol{x}$ in frame I_0, in subsequent frames I_n describe its trajectory along the sequence. We take these trajectory features taken into account in a next filtering stage comprising two steps.

First, for all pairs $\{I_0, I_n\}$, forward displacement fields $\boldsymbol{d}_{0,n}$ are spatiotemporally filtered considering spatiotemporal neighbouring forward vectors as well as the trajectories of spatial neighbouring pixels in the reference frame. To do so, the weights of our multilateral filter depend on spatial distance, colour similarity, matching cost and on a trajectory similarity that we introduce.

Until now, forward and backward displacement fields $\boldsymbol{d}_{0,n}$ and $\boldsymbol{d}_{n,0}$ have been estimated independently and carry complementary or contradictory information. In a second stage, they can be advantageously combined in a mutual refinement step. To this end, we present a joint multilateral filtering approach, both forward/backward and backward/forward.

Our experiments show that the optimal combination of short and long term matching does its job reducing the drift compared to optical-flow integration. Concerning temporary occlusions, while for single step methods (STEP1) it is impossible to estimate the trajectories of the occluded pixels after the occlusion (finally attaching all the tracks to the motion of the occluding object, which obliges to stop the trajectory), our multi-step fusion algorithm is able to circumvent the problem thanks to the long-step input displacement fields, as illustrated in Fig. 2.

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *PAMI*, 33(3):500–513, 2011.

[2] V. Lempitsky, S. Roth, and C. Rother. Fusionflow: Discrete-continuous optimization for optical flow estimation. In *CVPR*, 2008.

[3] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011.

[4] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[5] J. Wills, S. Agarwal, and S. Belongie. A feature-based approach for dense segmentation and estimation of large disparity motion. *IJCV*, 68(2):125–143, 2006.