Dense long-term motion estimation via Statistical Multi-Step Flow

Pierre-Henri Conze, Philippe Robert, Tomas Crivelli, Luce Morin

Dense long-term motion estimation via *Statistical Multi-Step Flow*

Pierre-Henri Conze[1,2], Philippe Robert[1], Tomás Crivelli[1] and Luce Morin[2]
[1]*Technicolor, Cesson-Sevigne, France*
[2]*INSA Rennes, IETR/UMR* 6164*, UEB, Rennes, France*
{*pierre-henri.conze, philippe.robert, tomas.crivelli*}*@technicolor.com, luce.morin@insa-rennes.fr*

Keywords: long-term motion estimation, dense point matching, statistical analysis, long-term trajectories, video editing

Abstract: We present *statistical multi-step flow*, a new approach for dense motion estimation in long video sequences. Towards this goal, we propose a two-step framework including an initial dense motion candidates generation and a new iterative motion refinement stage. The first step performs a combinatorial integration of elementary *optical flows* combined with a statistical candidate displacement fields selection and focuses especially on reducing motion inconsistency. In the second step, the initial estimates are iteratively refined considering several motion candidates including candidates obtained from neighboring frames. For this refinement task, we introduce a new energy formulation which relies on strong temporal smoothness constraints. Experiments compare the proposed statistical *multi-step* flow approach to state-of-the-art methods through both quantitative assessment using the *Flag* benchmark dataset and qualitative assessment in the context of video editing.

1 INTRODUCTION

Dense motion estimation has known significant improvements since early works but deals mainly with matching consecutive frames. Resulting dense motion fields, called *optical flows*, can straightforwardly be concatenated to describe the trajectories of each pixel along the sequence (Corpetti et al., 2002; Brox and Malik, 2010; Sundaram et al., 2010). However, both estimation and accumulation errors result in dense trajectories which can rapidly diverge and become inconsistent, especially for complex scenes including non-rigid deformations, large motion, zooming, poorly textured areas, illumination changes... Moreover, concatenating motion fields computed between consecutive frames does not allow to recover trajectories after temporary occlusions.

Recent works have contributed to the purpose of dense long-term motion estimation. Multi-frame *optical flow* formulations (Salgado and Sánchez, 2007; Papadakis et al., 2007; Werlberger et al., 2009; Volz et al., 2011) have been presented but their temporal smoothness constraints are generally limited to a small number of frames. (Sand and Teller, 2008) proposes a sophisticated framework to compute semi-dense trajectories using a particle representation but the full density is not achieved. To overcome these issues, Garg *et al.* describe in (Garg et al., 2013) a variational approach with subspace constraints to generate trajectories starting from a reference frame in a non-rigid context. They assume that the sequence of displacement of any point can be expressed as a linear combination of a low-rank motion basis. Therefore, trajectories are estimated assuming that they must lie close to this low dimensional subspace which implicitly acts as a long-term regularization. However, strong *a-priori* assumptions on scene contents must be provided and dense tracking of multiple objects is possible only if the reference frame is segmented.

The alternative concept of *multi-step* flow (Crivelli et al., 2012b; Crivelli et al., 2012a) focuses on how to construct dense fields of correspondences over extended time periods using *multi-step optical flows* (*optical flows* computed between consecutive frames or with larger inter-frame distances). *Multi-step* flow 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), *multi-step* flow 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*).

Despite its ability to handle both scenarios, *multi-step* flow has two main drawbacks. First, it performs the selection of displacement fields by relying only on classical *optical flow* assumptions that can sometimes 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 following frames of the sequence, until a new available *step* gives a chance to match with a correct location again.

These limitations can be resolved by extending to the whole sequence the combinatorial *multi-step* integration and the statistical selection described in (Conze et al., 2013) for dense motion estimation between a pair of distant frames. The underlying idea is to first consider a large set composed of combinations of *multi-step optical flows* and then to study the spatial redundancy of the resulting candidates through a statistical selection to finally select the best matches.

Toward our goal of dense motion estimation in long video shots, we present the *statistical multi-step flow* two-step framework. First, it extends (Conze et al., 2013) to generate several initial dense correspondences between the reference frame and each of the subsequent images independently. Second, we propose to provide an accurate final dense matching by applying a new iterative motion refinement which involves strong temporal smoothness constraints.

2 STATISTICAL *MULTI-STEP* FLOW

Let us consider a sequence of $N+1$ RGB images $\{I_n\}_{n\in[\![0,\ldots,N]\!]}$ including I_{ref} considered as a reference frame. In this work, we focus on dense motion estimation between the reference frame I_{ref} and each frame I_n of the sequence and we aim at computing *from-the-reference* and *to-the-reference* displacement fields. *From-the-reference* displacement fields link the reference frame I_{ref} to the other frames I_n and therefore describe the trajectory of each pixel of I_{ref} along the sequence. *To-the-reference* displacement fields connect each pixel of I_n to locations into I_{ref}.

The proposed *statistical multi-step flow* performs two main stages. The generation of several initial dense motion correspondences for each pair of frames $\{I_{ref}, I_n\}$ independently is described in Section 2.1. Section 2.2 presents the iterative motion refinement through strong temporal consistency constraints.

2.1 Initial motion candidates generation

The goal of the initial motion candidates generation is to compute for each pixel $\boldsymbol{x}_{ref}$ (resp. $\boldsymbol{x}_n$) of I_{ref} (resp. I_n) K candidate positions in I_n (resp. I_{ref}). Each pair of frames $\{I_{ref}, I_n\}$ is processed independently. Our explanations focus on the estimation of *from-the-reference* displacement fields. In the following, we describe the input data and recall the baseline method (Conze et al., 2013) before focusing on how it has been improved and extended to the whole sequence.

2.1.1 Input *optical flows* fields

As inputs, our method considers a set of *optical flow* fields estimated from each frame of the sequence including I_{ref}. These *optical flows* are previously estimated between consecutive frames or with larger

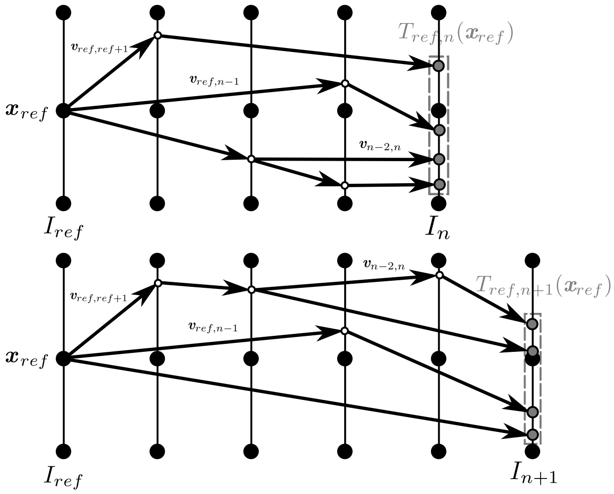


Figure 1: Multiple motion candidates are generated via a guided-random selection among all possible motion *paths*. This combinatorial integration (Conze et al., 2013) is done independently for each pair $\{I_{ref}, I_n\}$ which limits the correlation between candidates selected for neighbouring frames.

steps (Crivelli et al., 2012b), i.e. larger inter-frame distances. Let $S_n = \{s_1, s_2, \ldots, s_{Q_n}\} \subset \{1, \ldots, N-n\}$ be the set of Q_n possible *steps* at instant n. The following set of *optical flow* fields starting from I_n is therefore available: $\{\boldsymbol{v}_{n,n+s_1}, \boldsymbol{v}_{n,n+s_2}, \ldots, \boldsymbol{v}_{n,n+s_{Q_n}}\}$.

Input *optical flow* fields are provided with attached occlusion and inconsistency masks. For the pair $\{I_n, I_{n+s_i}\}$ with $s_i \in \{1, \ldots, N-n\}$, the occlusion mask attached to the *optical flow* field $\boldsymbol{v}_{n,n+s_i}$ indicates the visibility of each pixel of I_n in I_{n+s_i}. The inconsistency mask attached to $\boldsymbol{v}_{n,n+s_i}$ distinguishes consistent and inconsistent *optical flow* vectors among the ones starting from pixels marked as visible (Robert et al., 2012). This feature follows the idea that the *backward* flow should be the exact opposite of the *forward* flow.

2.1.2 Baseline method (Conze et al., 2013)

The combinatorial *multi-step* integration and the statistical selection on which we rely on work as follows.

For the current pair $\{I_{ref}, I_n\}$, the combinatorial *multi-step* integration consists in first of all considering all the possible *from-the-reference* motion *paths* which start from each pixel $\boldsymbol{x}_{ref}$, run through the sequence and end in I_n. These motion *paths* are built by concatenating all the possible sequences of un-occluded input *multi-step optical flow* vectors between I_{ref} and I_n. A reasonable number of N_s motion *paths* are then selected through limitations in terms of number of concatenations N_c and via a guided-random selection. Each remaining motion *path* leads to a candidate position in I_n (Fig. 1 *top*). Finally, we obtain a set $T_{ref,n}(\boldsymbol{x}_{ref}) = \{\boldsymbol{x}_n^i\}_{i\in[\![0,\ldots,K_{\boldsymbol{x}_{ref}}-1]\!]}$ of $K_{\boldsymbol{x}_{ref}}$ candidate positions in I_n for each pixel $\boldsymbol{x}_{ref}$ of I_{ref}.

A statistical-based selection stage then selects the optimal candidate position among $T_{ref,n}(\boldsymbol{x}_{ref})$. This procedure involves: 1) a statistical criterion which pre-selects a small set of candidates based on spatial density and intrinsic inconsistency values; 2) a global optimization which fuses these candidates to obtain the optimal one while including spatial regularization.

2.1.3 Improvements

The combinatorial *multi-step* integration and the statistical selection we briefly reviewed has been improved to provide further focus to inconsistency reduction between *from*/*to-the-reference* vectors. First, we use only *multi-step optical flow* vectors considered as consistent according to their inconsistency masks to generate motion *paths* between I_{ref} and I_n. Second, we introduce an outlier removal step before the statistical selection which orders the candidates of $T_{ref,n}(\boldsymbol{x}_{ref})$ with respect to their inconsistency values. A percentage $R_{\%}$ of bad candidates is removed and the selection is performed on the remaining ones. Third, at the end of the combinatorial integration and the selection procedure between I_{ref} and I_n, the optimal displacement field is incorporated into the processing between I_n and I_{ref} which aims at enforcing the motion consistency between *from*/*to-the-reference* fields.

Compared to (Conze et al., 2013), our displacement fields selection procedure combines differently statistical selection and global optimization. For each $\boldsymbol{x}_{ref} \in I_{ref}$, we select among $\boldsymbol{T}_{ref,n}(\boldsymbol{x}_{ref})$ $K_{sp} = 2 \times K$ candidates through statistical selection, with $K_{sp} < K_{\boldsymbol{x}_{ref}}$. Then, we randomly group by pairs these K_{sp} candidates and choose the K best ones $\overline{\boldsymbol{x}}_n^k$ $\forall k \in [\![0,\ldots,K-1]\!]$ by pair-wise fusing them following a global flow fusion approach. Finally, this same global optimization method fuses these K best candidates to obtain an optimal one: $\boldsymbol{x}_n^*$. In other words, these two last steps give a set of candidate displacement fields $\overline{\boldsymbol{d}}_{ref,n}^k$ and finally $\boldsymbol{d}_{ref,n}^*$, the optimal one. For pairs of frames relatively close or in case of temporary occlusions, the statistical selection is not adapted due to the small amount of candidates. Therefore, between $K+1$ and K_{sp} candidates, we use only the global optimization up to obtain the K best ones.

Our approach is applied bi-directionally. An exactly similar processing between I_n and I_{ref} leads to K initial *to-the-reference* candidate displacement fields.

2.1.4 Extention to the whole sequence

This improved version of the combinatorial integration and the statistical selection of (Conze et al., 2013) processes independently all the pairs $\{I_{ref}, I_n\}$. Only N_c, the maximum number of concatenations, changes

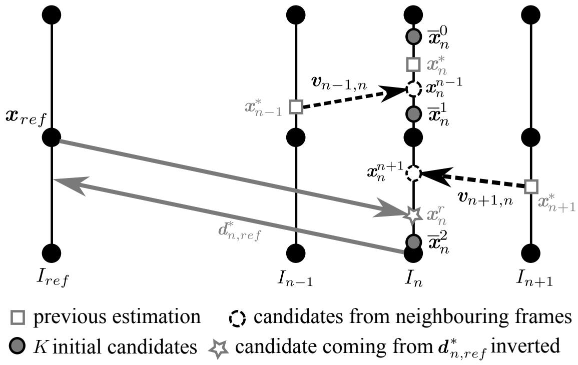


Figure 2: The displacement field $\boldsymbol{d}_{ref,n}^*$ is questionned by generating for each pixel $\boldsymbol{x}_{ref}$ competing candidates in I_n.

with respect to the temporal distance between frames. In practice, N_c is computed using Eq. (1) which leads to a good compromise between a too large number of concatenations which would lead to large propagation errors and the opposite situation which would limit the effectiveness of the statistical processing due to an insufficient number of candidates.

$$N_c(n) = \begin{cases} \mid n - ref \mid \text{ if } \mid n - ref \mid \leq 5 \\ \alpha_0 . \log 10(\alpha_1 . |n - ref|) \text{ otherwise} \end{cases} \quad (1)$$

The guided-random selection (Conze et al., 2013) which selects for each pair of frames $\{I_{ref}, I_n\}$ one part of all the possible motion *paths* limits the correlation between candidates respectively estimated for neighbouring frames. This avoids the situation in which a single estimation error is propagated and therefore badly influences the whole trajectory. The example Fig. 1 shows the motion *paths* selected by the guided-random selection for pairs $\{I_{ref}, I_n\}$ and $\{I_{ref}, I_{n+1}\}$. We notice that motion *paths* between I_{ref} and I_{n+1} are not highly correlated with those between I_{ref} and I_n. Indeed, the sets of *optical flow* vectors involved in both cases are not the same except for $\boldsymbol{v}_{ref,ref+1}$ and $\boldsymbol{v}_{ref,n-1}$ which are then concatenated with different vectors. $\boldsymbol{v}_{n-2,n}$ contributes for both cases but the considered vectors do not start from the same position. These considerations about the statistical independence of the resulting displacement fields are not addressed by existing methods for which a strong temporal correlation is inescapable.

2.2 Iterative motion refinement

The previous stage guarantees a low correlation between the initial motion candidates respectively estimated for pairs $\{I_{ref}, I_n\}$. Without losing this key characteristic, this second stage aims at iteratively refining the initial estimates while enforcing the temporal smoothness along the sequence.

We propose to question the matching between each pixel x_{ref} (resp. x_n) of I_{ref} (resp. I_n) and the selected position $\boldsymbol{x}_n^*$ (resp. $\boldsymbol{x}_{ref}^*$) in I_n (resp. I_{ref}) established during the previous iteration (or the initial motion candidates generation stage if the current iteration is the first one). For this task, we generate several competing candidates which are compared to $\boldsymbol{x}_n^*$ (resp. $\boldsymbol{x}_{ref}^*$) through a global optimization approach.

2.2.1 Competing candidates

The competing candidates used to question $\boldsymbol{x}_n^*$ (resp. $\boldsymbol{x}_{ref}^*$) are illustrated in Fig. 2 and deals with:

- the K initial candidate positions $\overline{\boldsymbol{x}}_n^k$ (resp. $\overline{\boldsymbol{x}}_{ref}^k$) $\forall k \in [\![0,\ldots,K-1]\!]$ (obtained Section 2.1),
- a candidate position coming from the previous estimation of $\boldsymbol{d}_{n,ref}^*$ (resp. $\boldsymbol{d}_{ref,n}^*$) which is inverted to obtain $\boldsymbol{x}_n^r$ (resp. $\boldsymbol{x}_{ref}^r$), as illustrated in Fig. 2,
- candidates from neighbouring frames to enforce temporal smoothing. Let W be the temporal window of width w centered around I_n. Between I_{ref} and I_n, we use the *optical flow* fields $\boldsymbol{v}_{m,n}$ between I_m and I_n with $m \in [\![n-\frac{w}{2},\ldots,n+\frac{w}{2}]\!]$ and $m \neq n$ to obtain from $\boldsymbol{x}_m^* \in I_m$ the new candidate $\boldsymbol{x}_n^m$ in I_n.

2.2.2 Global optimization approach

We perform a global optimization method in order to fuse the previously described competing candidates into a single optimal displacement field.

In the *from-the-reference* case, we introduce $L = \{l_{\boldsymbol{x}_{ref}}\}$ as a labeling of pixels $\boldsymbol{x}_{ref}$ where each label indicates $\boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}}$, one of the candidates listed above. Let $\boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}$ be the corresponding motion vector. We define the energy in Eq. (2) and minimize it with respect to L using *fusion moves* (Lempitsky et al., 2010):

$$E_{ref,n}(L) = E_{ref,n}^d(L) + E_{ref,n}^r(L) = \sum_{\boldsymbol{x}_{ref}} \rho_d(\varepsilon_{ref,n}^d) + \sum_{\boldsymbol{x}_{ref},\boldsymbol{y}_{ref}} \alpha_{\boldsymbol{x}_{ref},\boldsymbol{y}_{ref}} \rho_r(\left\| \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}(\boldsymbol{x}_{ref}) - \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{y}_{ref}}}(\boldsymbol{y}_{ref}) \right\|_1) \quad (2)$$

The data term $E_{ref,n}^d$, described with more details in Eq. (3), involves both matching cost and inconsistency value with respect to $\boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}$ (Conze et al., 2013). In addition, we propose to introduce strong temporal smoothness constraints into the energy formulation:

$$\varepsilon_{ref,n}^d = C(\boldsymbol{x}_{ref}, \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}(\boldsymbol{x}_{ref})) + Inc(\boldsymbol{x}_{ref}, \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}(\boldsymbol{x}_{ref})) + \sum_{\substack{m=n-\frac{w}{2} \\ m \neq n}}^{n+\frac{w}{2}} C(\boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}}, \boldsymbol{x}_m^* - \boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}}) + ed_{m,n} + ed_{n,m} \quad (3)$$

The temporal smoothness constraints translate into three new terms which are computed with respect to each neighbouring candidate $\boldsymbol{x}_m^*$ defined for the frames inside the temporal window W. These terms are illustrated in Fig. 3 and deal more precisely with:

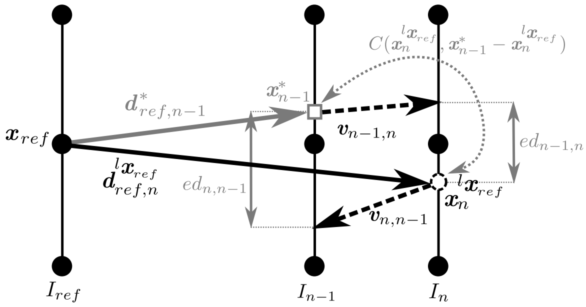


Figure 3: Matching cost and *Euclidean* distances $ed_{n,m}$ and $ed_{m,n}$ defined with respect to each temporal neighboring candidate $\boldsymbol{x}_m^*$ and involved in the proposed energy. These three terms act as strong temporal smoothness constraints.

- the matching cost between $\boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}} \in I_n$ and $\boldsymbol{x}_m^*$ of I_m,
- the *euclidean* distance $ed_{m,n}$ between $\boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}}$ and the ending point of the *optical flow* $\boldsymbol{v}_{m,n}$ starting from $\boldsymbol{x}_m^*$ (see Eq. (4)). $ed_{m,n}$ encourages the selection of $\boldsymbol{x}_n^m$, the candidate coming from I_m via the *optical flow* field $\boldsymbol{v}_{m,n}$ and therefore tends to strengthen the temporal smoothness. Indeed, for $\boldsymbol{x}_n^m$, the *euclidean* distance $ed_{m,n}$ is equal to 0.

$$ed_{m,n} = \left\| (\boldsymbol{x}_{ref} + \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}}) - (\boldsymbol{x}_{ref} + \boldsymbol{d}_{ref,m}^* + \boldsymbol{v}_{m,n}) \right\|_2 \quad (4)$$

- the *euclidean* distance $ed_{n,m}$ between $\boldsymbol{x}_m^*$ and the ending point of the *optical flow* vector $\boldsymbol{v}_{n,m}$ starting from $\boldsymbol{x}_n^{l_{\boldsymbol{x}_{ref}}}$ (see Eq. (5)). If $\boldsymbol{v}_{m,n}$ is consistent, i.e. $\boldsymbol{v}_{m,n} \approx -\boldsymbol{v}_{n,m}$, $ed_{n,m}$ is approximately equal to 0 for $\boldsymbol{x}_n^m$, the candidate coming from I_m, whose selection is again promoted.

$$ed_{n,m} = \left\| (\boldsymbol{x}_{ref} + \boldsymbol{d}_{ref,m}^*) - (\boldsymbol{x}_{ref} + \boldsymbol{d}_{ref,n}^{l_{\boldsymbol{x}_{ref}}} + \boldsymbol{v}_{n,m}) \right\|_2 \quad (5)$$

The regularization term $E_{ref,n}^r$ involves motion similarities with neighbouring positions, as shown in Eq. (2). $\alpha_{\boldsymbol{x}_{ref},\boldsymbol{y}_{ref}}$ accounts for local color similarities in the reference frame I_{ref}. The robust functions ρ_d and ρ_r are respectively the negative log of a *Student-t* distribution and the *Geman-McClure* function.

The refinement of *to-the-reference* displacement fields with our approach is straightforward except that the data term involves neither the matching cost between the current candidate and the temporal neighbouring one nor the *euclidean* distance $ed_{m,n}$ due to trajectories which can not be handled in this direction.

The global optimization method fuses the displacement fields by pairs and finally chooses to update or not the previous estimations with one of the previously described candidates. The motion refinement phase consists in applying this technique for each pair of frames $\{I_{ref}, I_n\}$ in *from-the-reference* and *to-the-reference* directions. The pairs $\{I_{ref}, I_n\}$ are processed in a random order in order to encourage temporal smoothness without introducing a sequential correlation between the resulting displacement fields.

This motion refinement phase is repeated iteratively N_{it} times where one iteration corresponds to the processing of all the pairs $\{I_{ref}, I_n\}$. The proposed *statistical multi-step flow* is done once the initial motion candidates generation and the N_{it} iterations of motion refinement have been performed.

3 EXPERIMENTS

Our experiments focus on the following sequences: *MPI S1* (Granados et al., 2012) Fig.4 and 6*a-h*, *Hope* Fig.6*i-p*, *Newspaper* Fig.6*q-t*, *Walking Couple* Fig.7 and *Flag* (Garg et al., 2013) Fig.8. The proposed *statistical multi-step flow* is referred to as *StatFlow* in the following. For the experiments, the following parameters have been used: $N_c = 7$, $N_s = 100$, $R_{\%} = 50\%$, $K = 3$, $\alpha_0 = 3$, $\alpha_1 = 15$, $w = 5$. The set of *steps* and input *optical flow* estimators will be specified for each experiment and each sequence.

Experiments have been conducted as follows. In Section 3.1, we evaluate the performance of our extended version of the combinatorial integration and the statistical selection (Conze et al., 2013) through registration and PSNR assessment. The effects of the iterative motion refinement are also studied. Then, we compare *StatFlow* to state-of-the-art methods through quantitative assessment using the *Flag* dataset (Garg et al., 2013) (Section 3.2) and qualitative assessment via texture propagation and tracking (Section 3.3).

3.1 Registration and PSNR assessment

The first experiment aims at showing how the improvements we made with respect to (Conze et al., 2013) impacts the quality of the displacement fields. We focus on frames pairs taken from *MPI S1* and *Newspaper* (*NP*). The sets of *steps* are $1-5$, 10 (*NP*), 15 (*MPI S1*), 20 (*NP*) and 30 (*NP*). The algorithms are performed taking input *multi-step optical flows* computed with a 2D version of the disparity estimator described in (Robert et al., 2012), referred to as *2D-DE*.

We compare the optimal displacement fields obtained in output of our initial motion estimates generation (Section 2.1) with those resulting from (Conze et al., 2013). The comparison is done through registration and PSNR assessment. For a given pair $\{I_{ref}, I_n\}$, the final fields are used to reconstruct I_{ref} from I_n through motion compensation and color

Frame pairs	{25,45}	{25,46}	{25,47}	{25,48}
(Conze et al., 2013)	21.83	24.98	25.56	25.83
StatFlow initial phase	**29.02**	**28.4**	**27.27**	**27.23**
Frame pairs	{25,49}	{25,50}	{25,51}	{25,52}
(Conze et al., 2013)	25.04	24.83	24.48	24.3
StatFlow initial phase	**26.84**	**26.33**	**26.1**	**25.69**

Table 1: Registration and PSNR assessment with the combinatorial integration and the statistical selection introduced in (Conze et al., 2013) and the proposed extended version described in Section 2.1 (initial phase of *StatFlow*). PSNR scores are computed on the kiosk of *MPI S1* (Fig. 4).

Frame pairs	{160,180}	{160,190}	{160,200}
(Conze et al., 2013)	22.50	21.21	18.59
StatFlow initial phase	22.70	21.39	19.28
StatFlow	**22.93**	**22.18**	**20.25**
Frame pairs	{160,210}	{160,220}	{160,230}
(Conze et al., 2013)	17.12	15.87	15.76
StatFlow initial phase	18.21	17.12	16.58
StatFlow	**18.68**	**17.40**	**16.81**

Table 2: Registration and PSNR assessment with: 1) combinatorial integration and statistical selection introduced in (Conze et al., 2013), 2) proposed extended version (*StatFlow* init. phase), 3) whole *StatFlow* method. PSNR scores are computed on whole images of *Newspaper* (Fig.6*q-t*).

PSNR scores are computed between I_{ref} and the registered frame for non-occluded pixels.

Tables 1 and 2 show the PSNR scores for various distances between I_{ref} and I_n respectively on the kiosk of *MPI S1* (Fig.4) and on whole images of *Newspaper* (Fig.6*q-t*). Results on *MPI S1* show that the initial phase of *StatFlow* outperforms the combinatorial integration and the statistical selection of (Conze et al., 2013) for all pairs. An example of registration of the kiosk for a distance of 20 frames is given Fig.4. *Multi-step* estimations deal satisfactorily with the temporary occlusion. Experiments on *Newspaper* reveal the same finding: the novelty in terms of inconsistency reduction improves the displacement fields quality. Moreover, the iterative motion refinement stage ($N_{it} = 9$) allows to obtain better PSNR scores for all pairs compared to the initial stage of *StatFlow*.

3.2 Comparisons with *Flag* dataset

Quantitative results have been obtained using the dense ground-truth *optical flow* data provided by the *Flag* dataset (Garg et al., 2013) for the *Flag* sequence (Fig. 8). Experiments focus on:

- direct estimation between each pair $\{I_{ref}, I_n\}$ using *LDOF* (Brox and Malik, 2011), *ITV-L1* (Wedel et al., 2009) and the keypoint-based non-rigid registration of (Pizarro and Bartoli, 2012),
- concatenation of *optical flows* computed between consecutive frames using *LDOF* (*LDOF acc*),

(a) I_{25} (b) I_{40} (c) I_{45} (d) I_{25} (e) (Conze et al., 2013) (f) *StatFlow* initial phase

Figure 4: Source frames of the *MPI S1* sequence (Granados et al., 2012) and reconstruction of the kiosk of I_{25} from I_{45} with: e) the combinatorial integration and the statistical selection introduced in (Conze et al., 2013), f) the proposed extended version described in Section 2.1 (initial phase of *StatFlow*). Black boxes focus on differences between both methods.

Method	RMS endpoint error (pixels)
***StatFlow* (*LDOF*)**	**0.69**
MSF (Crivelli et al., 2012a) (*LDOF*)	1.41
LDOF direct (Brox and Malik, 2011)	1.74
LDOF acc (Brox and Malik, 2011)	4
MFSF-PCA (Garg et al., 2013)	**0.69**
MFSF-DCT (Garg et al., 2013)	0.80
(Pizarro and Bartoli, 2012) *direct*	1.24
ITV-L1 direct (Wedel et al., 2009)	1.43

Table 3: RMS endpoint errors for different methods on the *Flag* benchmark dataset (Garg et al., 2013).

- *multi-frame subspace flow* (*MFSF*) (Garg et al., 2013) using PCA or DCT basis,
- *multi-step* flow fusion (*MSF*) (Crivelli et al., 2012a) with *LDOF multi-step optical flows*,
- *StatFlow* ($N_{it} = 3$) with *LDOF optical flows*.

For the comparison task, Tab. 3 gives for all the previously described methods the RMS (*root mean square*) endpoint errors between the respective obtained displacement fields and the ground-truth data. RMS errors are estimated for all the foreground pixels and for all the pairs of frames $\{I_{ref}, I_n\}$ together. RMS errors computed for each pair of frames are shown in Fig.5 for all the methods based on *LDOF*: *LDOF direct*, *LDOF acc*, *MSF* (*LDOF*) and *StatFlow* (*LDOF*). The last two *multi-step* strategies have considered as inputs steps $1-5$, 8, 10, 15, 20, 25, 30, 40 and 50.

We can firstly observe that *LDOF acc* rapidly diverge. This is due to both estimation errors which are propagated along trajectories and accumulation errors inherent to the interpolation process. Moreover, the results obtained through direct motion estimation are reasonably good, especially for (Pizarro and Bartoli, 2012). *LDOF direct* gives a lower RMS endpoint error than *LDOF acc* (1.74 against 4). However, it is not possible to draw conclusions in the light of the *Flag* sequence because the flag comes back approximately to its initial position at the end of the sequence (Fig.8*a*,*g*). Motion estimation for complex scenes cannot generally rely only on direct estimation and combining *optical flow* accumulations and direct matching is clearly a more suitable strategy.

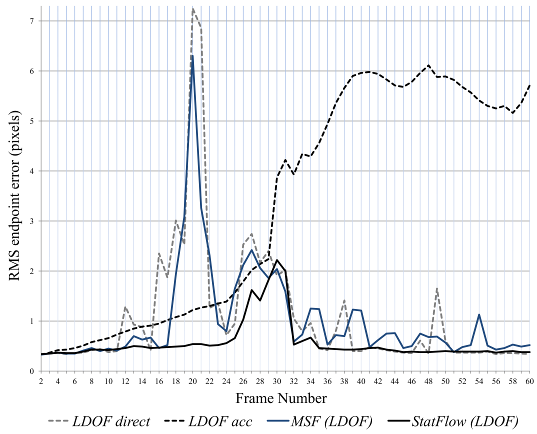


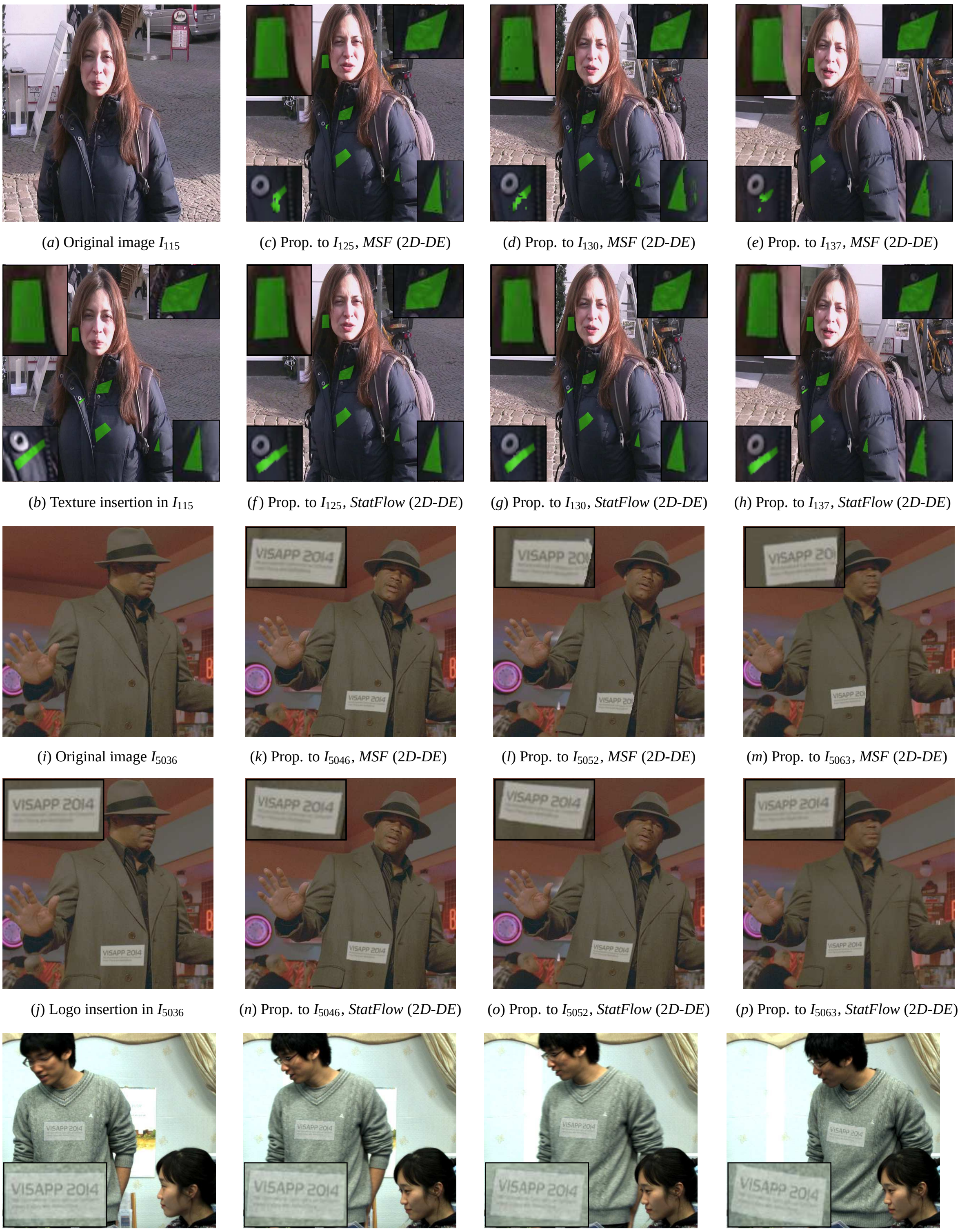
Figure 5: RMS endpoint errors for each pair $\{I_{ref}, I_n\}$ along *Flag* sequence (Fig. 8) with different methods.

Tab. 3 and Fig. 5 prove that with the same *optical flows* as inputs, *StatFlow* shows a clear improvement compared to *MSF* (0.69 against 1.41). Although both methods achieve the same quality for first pairs or for some pairs which coincide with existing *steps*, other displacement fields are computed with a better accuracy using *StatFlow*. Moreover, *StatFlow*(*LDOF*) reaches the same RMS error with respect to *MFSF-PCA*, the best one of the *MFSF* approaches, with 0.69. This proves that *StatFlow* is competitive compared to challenging state-of-the-art methods.

3.3 Texture propagation and tracking

We aim now at showing that our method provides satisfying results in a wide set of complex scenes. Moreover, we focus on the comparison between *StatFlow* ($N_{it} = 9$) and *MSF* (Crivelli et al., 2012a) to prove that *StatFlow* performs a more efficient integration and selection procedure compared to *MSF* using the same *optical flows* as inputs. Experiments have been firstly conducted in the context of video editing: we evaluate the accuracy of both methods by motion compensating in I_n $\forall n$ textures/logos manually inserted in I_{ref}.

In Fig. 6 and 7, textures/logos have been respectively inserted in I_{115} of *MPI S*1, I_{5036} of *Hope*, I_{230} of *Newspaper* and I_0 of *Walking Couple*. *To-the-reference* fields computed with *StatFlow (*2*D-DE)* and *MSF (*2*D-DE)* serve to propagate textures/logos up to respectively I_{137}, I_{5063}, I_{170} and I_{40}. 2*D-DE* has been



(a) Original image I_{115} (c) Prop. to I_{125}, *MSF* (*2D-DE*) (d) Prop. to I_{130}, *MSF* (*2D-DE*) (e) Prop. to I_{137}, *MSF* (*2D-DE*)

(b) Texture insertion in I_{115} (f) Prop. to I_{125}, *StatFlow* (*2D-DE*) (g) Prop. to I_{130}, *StatFlow* (*2D-DE*) (h) Prop. to I_{137}, *StatFlow* (*2D-DE*)

(i) Original image I_{5036} (k) Prop. to I_{5046}, *MSF* (*2D-DE*) (l) Prop. to I_{5052}, *MSF* (*2D-DE*) (m) Prop. to I_{5063}, *MSF* (*2D-DE*)

(j) Logo insertion in I_{5036} (n) Prop. to I_{5046}, *StatFlow* (*2D-DE*) (o) Prop. to I_{5052}, *StatFlow* (*2D-DE*) (p) Prop. to I_{5063}, *StatFlow* (*2D-DE*)

(q) Logo insertion in I_{230} (r) Prop. to I_{210}, *StatFlow* (*2D-DE*) (s) Prop. to I_{196}, *StatFlow* (*2D-DE*) (t) Prop. to I_{170}, *StatFlow* (*2D-DE*)

Figure 6: Texture/logo insertion in I_{115} (resp. I_{5036} and I_{230}) and propagation along the *MPI-S*1 (resp. *Hope* and *Newspaper*) sequence up to I_{137} (resp. I_{5063} and I_{170}) using: 1) *multi-step* flow fusion (*MSF*) (Crivelli et al., 2012a) with *multi-step optical flow* fields from (Robert et al., 2012) (*2D-DE*): *MSF*(*2D-DE*); 2) the proposed *statistical multi-step flow* (*StatFlow*) with *2D-DE multi-step optical flow* fields: *StatFlow* (*2D-DE*).

(a) Original image I_0 (d) Propagation to I_{20}, *LDOF acc* (e) Propagation to I_{25}, *LDOF acc* (f) Propagation to I_{40}, *LDOF acc*

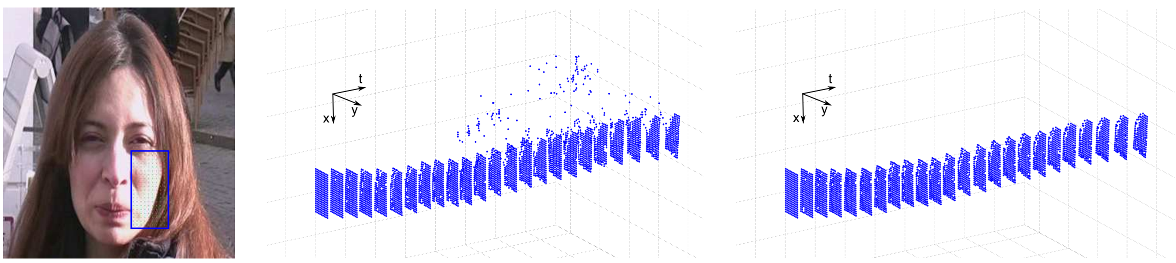
(b) Logo insertion in I_0 (g) Prop. to I_{20}, *MSF* (*2D-DE*) (h) Prop. to I_{25}, *MSF* (*2D-DE*) (i) Prop. to I_{40}, *MSF* (*2D-DE*)

(c) Original image I_{40} (j) Prop. to I_{20}, *StatFlow* (*2D-DE*) (k) Prop. to I_{25}, *StatFlow* (*2D-DE*) (l) Prop. to I_{40}, *StatFlow* (*2D-DE*)

Figure 7: Texture insertion in I_0 and propagation up to I_{40} (*Walking Couple* sequence). We compare: d-f) concatenation of *LDOF* (Brox and Malik, 2011) *optical flow* fields computed between consecutive frames (*LFOF acc*); g-i) *multi-step* flow fusion (*MSF*) (Crivelli et al., 2012a) using *multi-step optical flow* fields from (Robert et al., 2012) (*2D-DE*); j-l) the proposed *statistical multi-step flow* (*StatFlow*) using *2D-DE multi-step optical flow* fields.

(a) I_1 (b) I_{10} (c) I_{20} (d) I_{30} (e) I_{40} (f) I_{50} (g) I_{60}

Figure 8: Source frames of the *Flag* sequence (Garg et al., 2013).



(a) I_{115} with tracking area (b) Point tracking from I_{115} to I_{138}, *MSF* (*2D-DE*) (c) Point tracking from I_{115} to I_{138}, *StatFlow* (*2D-DE*)

Figure 9: Point tracking from I_{115} up to I_{138}, *MPI-S1* sequence (Granados et al., 2012). We compare: b) *multi-step* flow fusion (*MSF*) (Crivelli et al., 2012a) using *multi-step optical flow* fields from (Robert et al., 2012) (*2D-DE*); c) the proposed *statistical multi-step flow* (*StatFlow*) method using *2D-DE multi-step optical flow* fields.

chosen for its good results for video editing tasks. The *steps* involved are: $1-5$, 8 (*Hope*), 10, 15 (except for *NP*), 20 (*Hope*, *NP*), 30 (*MPI S*1, *NP*).

Given these results, it appears that *MSF* sometimes distorts structures (bottom left zoom Fig.6*c*-*e*, Fig.6*l*,*m*), makes shadow textures appear (bottom right zoom Fig.6*c*-*e*) and does not estimate motion with accuracy (top right zoom Fig.6*e*, Fig.6*l*,*m*). Visual results with *StatFlow* reveal a better long-term propagation (see also Fig.6*r*-*t*). Fig.7 compares *StatFlow(2D-DE)* and *MSF(2D-DE)* with *LDOF acc*. We observe that *LDOF acc* badly performs motion estimation for periodic structures. *MSF* encounters also matching issues (Fig.7*h*) whereas *StatFlow* performs propagation without any visible artifacts.

Finally, *StatFlow* and *MSF* are assessed through point tracking. In Fig. 9, the bottom right part of the woman face is tracked from I_{115} to I_{138} (*MPI S*1). The 2*D*+*t* visualization indicates that some trajectories drift to the background with *MSF*. This illustrates the inherent issue of *MSF* which propagates estimation errors due to the sequential processing. Conversely, *StatFlow* provides accurate fields while limiting the temporal correlation between displacement fields respectively estimated for neighbouring frames.

4 CONCLUSION

We present *statistical multi-step flow*, a two-step framework which performs dense long-term motion estimation. Our method starts by generating initial dense correspondences with a focus on inconsistency reduction. For this task, we perform a combinatorial integration of consistent *optical flows* followed by an efficient statistical selection. This procedure is applied independently between a reference frame and each frame of the sequence. It guarantees a low temporal correlation between the resulting correspondences respectively estimated for each of these pairs. We propose then to enforce temporal smoothness through a new iterative motion refinement. It considers several motion candidates including candidates from neighboring frames and involves a new energy formulation with temporal smoothness constraints. Experiments evaluate the effectiveness of our approach compared to state-of-the-art methods through quantitative assessment using dense ground-truth data and qualitative assessment via texture propagation and tracking for a wide set of complex scenes.

REFERENCES

Brox, T. and Malik, J. (2010). Object segmentation by long term analysis of point trajectories. *European Conference on Computer Vision*, pages 282–295.

Brox, T. and Malik, J. (2011). Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513.

Conze, P.-H., Crivelli, T., Robert, P., and Morin, L. (2013). Dense motion estimation between distant frames: combinatorial *multi-step* integration and statistical selection. In *IEEE International Conference on Image Processing*.

Corpetti, T., Mémin, É., and Pérez, P. (2002). Dense estimation of fluid flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):365–380.

Crivelli, T., Conze, P.-H., Robert, P., Fradet, M., and Pérez, P. (2012a). *Multi-Step Flow Fusion*: Towards accurate and dense correspondences in long video shots. *British Machine Vision Conference*.

Crivelli, T., Conze, P.-H., Robert, P., and Pérez, P. (2012b). From optical flow to dense long term correspondences. In *IEEE International Conference on Image Processing*.

Garg, R., Roussos, A., and Agapito, L. (2013). A variational approach to video registration with subspace constraints. *International Journal of Computer Vision*.

Granados, M., Kim, K. I., Tompkin, J., Kautz, J., and Theobalt, C. (2012). MPI-S1. http://www.mpi-inf.mpg.de/ granados/projects/vidbginp/index.html.

Lempitsky, V., Rother, C., Roth, S., and Blake, A. (2010). Fusion moves for *Markov* random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1392–1405.

Papadakis, N., Corpetti, T., and Mémin, E. (2007). Dynamically consistent optical flow estimation. In *IEEE International Conference on Computer Vision*.

Pizarro, D. and Bartoli, A. (2012). Feature-based deformable surface detection with self-occlusion reasoning. *International Journal of Computer Vision*.

Robert, P., Thébault, C., Drazic, V., and Conze, P.-H. (2012). Disparity-compensated view synthesis for s3D content correction. In *SPIE IS&T Electronic Imaging Stereoscopic Displays and Applications*.

Salgado, A. and Sánchez, J. (2007). Temporal constraints in large optical flow estimation. In *Computer Aided Systems Theory Eurocast*, pages 709–716.

Sand, P. and Teller, S. J. (2008). Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 80(1):72–91.

Sundaram, N., Brox, T., and Keutzer, K. (2010). Dense point trajectories by GPU-accelerated large displacement *optical flow*. *European Conference on Computer Vision*, pages 438–451.

Volz, S., Bruhn, A., Valgaerts, L., and Zimmer, H. (2011). Modeling temporal coherence for *optical flow*. In *IEEE International Conference on Computer Vision*.

Wedel, A., Pock, T., Zach, C., Bischof, H., and Cremers, D. (2009). An improved algorithm for *TV-L1 optical flow*. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, pages 23–45. Springer.

Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., and Bischof, H. (2009). Anisotropic *Huber-L1 optical flow*. *British Machine Vision Conference*.