Dense long-term motion estimation via Statistical Multi-Step Flow
Pierre-Henri Conze, Philippe Robert, Tomas Crivelli, Luce Morin

To cite this version:


HAL Id: hal-00910831
https://hal.archives-ouvertes.fr/hal-00910831
Submitted on 3 Dec 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Dense long-term motion estimation via \textit{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 \textit{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 \textit{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 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 \textit{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 \textit{multi-step optical flows} (optical flows computed between consecutive frames or with larger inter-frame distances). \textit{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 \textit{multi-step optical flows} with displacement vectors already computed for neighbouring frames. \textit{Multi-step} estimations can handle temporary occlusions since they can \textit{jump} occluding objects. Contrary to (Garg et al., 2013), \textit{multi-step} flow considers both trajectory estimation between a reference frame and all the images of the sequence (\textit{from-the-reference}) and motion estimation to match each image to the reference frame (\textit{to-the-reference}).

Despite its ability to handle both scenarios, \textit{multi-step} flow has two main drawbacks. First, it performs the selection of displacement fields by relying only on classical \textit{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 \textit{step} gives a chance to match with a correct location again.

These limitations can be resolved by extending to the whole sequence the combinatorial \textit{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 \( \{ I_{n} \}_{n=0}^{N} \) 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 \( x_{ref} \) (resp. \( x_{n} \)) 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 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: \( \{ v_{n,n+s_{1}}, v_{n,n+s_{2}}, \ldots, v_{n,n+s_{Q_{n}}} \} \).

Input optical flow fields are provided with attached occlusion and inconsistency masks. For the pair \( \{ I_{n}, I_{n+s} \} \) with \( s_{i} \in \{ 1, \ldots, N-n \} \), the occlusion mask attached to the optical flow field \( v_{n,n+s} \) indicates the visibility of each pixel of \( I_{n} \) in \( I_{n+s} \). The inconsistency mask attached to \( v_{n,n+s} \) 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 \( x_{ref} \), 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 \( N_{c} \) 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 \( T_{ref,n}(x_{ref}) = \{ x_{i}^{t} \}_{t=n+1}^{n+N_{c}-1} \) of \( K_{x_{ref}} \) candidate positions in \( I_{n} \) for each pixel \( x_{ref} \) of \( I_{ref} \).

![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.](image-url)
A statistical-based selection stage then selects the optimal candidate position among \( \text{I}_{\text{ref},n}(x_{\text{ref}}) \). 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 \( \text{I}_{\text{ref}} \) and \( \text{I}_n \). Second, we introduce an outlier removal step before the statistical selection which orders the candidates of \( \text{T}_{\text{ref},n}(x_{\text{ref}}) \) with respect to their inconsistency values. A percentage \( R_b \) 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 \( \text{I}_{\text{ref}} \) and \( \text{I}_n \), the optimal displacement field is incorporated into the processing between \( \text{I}_n \) and \( \text{I}_{\text{ref}} \) 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 \( x_{\text{ref}} \in \text{I}_{\text{ref}} \), we select among \( \text{T}_{\text{ref},n}(x_{\text{ref}}) \) \( K_{sp} = 2 \times K \) candidates through statistical selection, with \( K_{sp} < K_{x_{\text{ref}}} \). Then, we randomly group by pairs these \( K_{sp} \) candidates and choose the \( K \) best ones \( \text{I}_{\text{ref}}(x_{\text{ref}}) \) with respect to their inconsistency values. A percentage \( R_b \) 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 \( \text{I}_{\text{ref}} \) and \( \text{I}_n \), the optimal displacement field is incorporated into the processing between \( \text{I}_n \) and \( \text{I}_{\text{ref}} \) 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 \( x_{\text{ref}} \in \text{I}_{\text{ref}} \), we select among \( \text{T}_{\text{ref},n}(x_{\text{ref}}) \) \( K_{sp} = 2 \times K \) candidates through statistical selection, with \( K_{sp} < K_{x_{\text{ref}}} \). Then, we randomly group by pairs these \( K_{sp} \) candidates and choose the \( K \) best ones \( \text{I}_{\text{ref}}(x_{\text{ref}}) \) with respect to their inconsistency values. A percentage \( R_b \) 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 \( \text{I}_{\text{ref}} \) and \( \text{I}_n \), the optimal displacement field is incorporated into the processing between \( \text{I}_n \) and \( \text{I}_{\text{ref}} \) which aims at enforcing the motion consistency between from/to-the-reference fields.

Our approach is applied bi-directionally. An exactly similar processing between \( \text{I}_n \) and \( \text{I}_{\text{ref}} \) leads to \( K \) initial to-the-reference candidate displacement fields.

### 2.1.4 Extension 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 \{\( \text{I}_{\text{ref}}, \text{I}_n \)\}. Only \( N_c \), the maximum number of concatenations, changes

---

**Figure 2:** The displacement field \( d_{\text{ref},n}^e \) is questioned by generating for each pixel \( x_{\text{ref}} \) competing candidates in \( \text{I}_n \), 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.

\[
N_c(n) = \begin{cases} 
|n - \text{ref}| \text{ if } |n - \text{ref}| \leq 5 \\
\alpha_0 \log 10(|\alpha_1|n - \text{ref})) \text{ otherwise}
\end{cases}
\]

The guided-random selection (Conze et al., 2013) which selects for each pair of frames \{\( \text{I}_{\text{ref}}, \text{I}_n \)\} 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 \{\( \text{I}_{\text{ref}}, \text{I}_n \)\} and \{\( \text{I}_{\text{ref}}, \text{I}_{n+1} \)\}. We notice that motion paths between \( \text{I}_{\text{ref}} \) and \( \text{I}_{n+1} \) are not highly correlated with those between \( \text{I}_{\text{ref}} \) and \( \text{I}_n \). Indeed, the sets of optical flow vectors involved in both cases are not the same except for \( v_{\text{ref},n+1} \) and \( v_{\text{ref},n+1} \) which are then concatenated with different vectors. \( v_{n-2} \) 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 \{\( \text{I}_{\text{ref}}, \text{I}_n \)\}. 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 \( x_{\text{ref}} \) (resp. \( x_n \)) of \( \text{I}_{\text{ref}} \) (resp. \( \text{I}_n \)) and the
selected position $x_m^*(\text{resp. } x_{m}^\text{ref})$ 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 $x_m^*$ (resp. $x_{m}^\text{ref}$) through a global optimization approach.

### 2.2.1 Competing candidates

The competing candidates used to question $x_m^*$ (resp. $x_{m}^\text{ref}$) are illustrated in Fig. 2 and deals with:

- the $K$ initial candidate positions $X_k^0$ (resp. $X_k^\text{ref}$) \( \forall k \in \{0, \ldots, K-1\} \) (obtained Section 2.1),
- a candidate position coming from the previous estimation of $d_{n,ref}^*$ (resp. $d_{ref,n}^*$) which is inverted to obtain $x_{n,m}^*$ (resp. $x_{ref,n}^*$), 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 $v_{n,m}$ between $I_m$ and $I_n$ with $m \in \{n-w, \ldots, n+w\}$ and $m \neq n$ to obtain from $x_{n,m}^*$ in $I_m$ the new candidate $x_n^*$ 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 $L = \{I_{ref}\}$ as a labeling of pixels $x_{ref}$ where each label indicates $x_{n}^\text{ref}$, one of the candidates listed above. Let $d_{ref,n}^{ref}$ 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):

$$
E_{ref,n}(L) = E_{ref,n}^d(L) + E_{ref,n}^e(L) = \sum_{x_{ref}} \rho_d(d_{ref,n}^{ref}(x_{ref})) + \sum_{x_{ref} \notin \{x_{ref}\}} \alpha_{x_{ref},y_{ref}} \rho_r(\|d_{ref,n}^{ref}(x_{ref}) - d_{ref,n}^{ref}(y_{ref})\|_2)
$$

(2)

The data term $E^d_{ref,n} \text{ and the energy formulation:} \nabla x_n^\text{ref} = C(x_{ref},d_{ref,n}^{ref}(x_{ref}))+\text{Inc}(x_{ref},d_{ref,n}^{ref}(x_{ref})) + \sum_{m-m}^{n+w} C(x_{n,m},x_m^*-x_{n,m}^*), \text{ed}_{n,m} + \text{ed}_{m,n}

(3)

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

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

$$
ed_{n,m} = \left\| (x_{ref} + d_{ref,n}^{ref}) - (x_{ref} + d_{ref,m}^{ref} + v_{n,m}) \right\|_2 \tag{4}
$$

- the euclidean distance $ed_{n,m}$ between $x_{n}^*$ and the ending point of the optical flow vector $v_{n,m}$ starting from $x_{n,m}^*$ (see Eq. (5)). $ed_{n,m}$ accounts for local color similarities in the reference frame $I_{ref}$. $\alpha_{x_{ref},y_{ref}}$, the candidate coming from $I_m$, whose selection is again promoted.

$$
ed_{n,m} = \left\| (x_{ref} + d_{ref,n}^{ref}) - (x_{ref} + d_{ref,m}^{ref} + v_{n,m}) \right\|_2 \tag{5}
$$

The regularization term $E^r_{ref,n}$ involves motion similarities with neighbouring positions, as shown in Eq. (2). $\alpha_{x_{ref},y_{ref}}$ accounts for local color similarities in the reference frame $I_{ref}$. The robust functions $\rho_d$ and $\rho_r$ 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 $ed_{n,m}$ 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 \( N_t \) 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 \( N_t \) 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 6a-h, Hope Fig.6i-p, Newspaper Fig.6q-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: \( N_t = 7 \), \( N_c = 100 \), \( R_{c0} = 50\% \), \( K = 3 \), \( \alpha_0 = 3 \), \( \alpha_1 = 15 \), \( w = 5 \). 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 registration.

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

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

- **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** ($N_f = 3$) 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 $\{I_{ref}, I_t\}$ 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.8a,g). Motion estimation for complex scenes cannot generally rely on direct estimation and combining optical flow accumulations and direct matching is clearly a more suitable strategy.

![Image 69x650 to 123x746](image69x650to123x746)

Figure 4: Source frames of the MPI S1 sequence (Granados et al., 2012) and reconstruction of the kiosk of I_{25} from I_{25} 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.

![Image 156x692 to 213x746](image156x692to213x746)

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

![Image 309x692 to 365x747](image309x692to365x747)

Figure 5: RMS endpoint errors for each pair $\{I_{ref}, I_t\}$ 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 ($N_f = 9$) 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 $I_t \forall n$ textures/logos manually inserted in $I_{ref}$.

In Fig. 6 and 7, textures/logos have been respectively inserted in $I_{115}$ of MPI S1, $I_{90}$ of Hope, $I_{230}$ of Newspaper and $I_0$ of Walking Couple. To-the-reference fields computed with StatFlow (2D-DE) and MSF (2D-DE) serve to propagate textures/logos up to respectively $I_{137}$, $I_{90}$, $I_{70}$ and $I_0$. 2D-DE has been
Figure 6: Texture/logo insertion in $I_{115}$ (resp. $I_{5036}$ and $I_{230}$) and propagation along the MPI-S1 (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).
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 (LDOF 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.

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

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: $1 \rightarrow 5, 8$ (Hope), $10, 15$ (except for NP), $20$ (Hope, NP), $30$ (MPI S1, NP).

Given these results, it appears that MSF sometimes distorts structures (bottom left zoom Fig.6c-e, Fig.6.l.m), makes shadow textures appear (bottom right zoom Fig.6c-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.6r-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.7h) 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 S1). The 2D+1 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


Disparity-compensated view synthesis for 3D content correction. In SPIE IS&T Electronic Imaging Stereoscopic Displays and Applications.


Modeling temporal coherence for optical flow. In IEEE International Conference on Computer Vision.
