Dense motion estimation between distant frames: combinatorial multi-step integration and statistical selection

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

To cite this version:


HAL Id: hal-00868251
https://hal.science/hal-00868251
Submitted on 1 Oct 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 MOTION ESTIMATION BETWEEN DISTANT FRAMES: COMBINATORIAL MULTI-STEP INTEGRATION AND STATISTICAL SELECTION

Pierre-Henri Conze∗† Tomás Crivelli∗ Philippe Robert∗ Luce Morin†

∗Technicolor
†INSA Rennes, IETR/UMR 6164, UEB

ABSTRACT

Accurate estimation of dense point correspondences between two distant frames of a video sequence is a challenging task. To address this problem, we present a combinatorial multi-step integration procedure which allows one to obtain a large set of candidate motion fields between the two distant frames by considering multiple motion paths across the video sequence. Given this large candidate set, we propose to perform the optimal motion vector selection by combining a global optimization stage with a new statistical processing. Instead of considering a selection only based on intrinsic motion field quality and spatial regularization, the statistical processing exploits the spatial distribution of candidates and introduces an intra-candidate quality based on forward-backward consistency. Experiments evaluate the effectiveness of our method for distant motion estimation in the context of video editing.

Index Terms— motion estimation, statistical analysis, dense point matching, distant frames

1. INTRODUCTION

Despite rapid and significant progress since early formulations [1, 2], optical flow estimation still remains an open issue with crucial implications in computer vision. State-of-the-art methods [3, 4, 5, 6, 7, 8, 9] have shown to be highly accurate for estimating dense motion fields between consecutive frames of a video sequence. However, they show limitations when applied to distant frames. The classical optical flow assumptions are not verified in this case, especially for difficult situations such as illumination changes, large motion, temporal occlusions, zooming, non-rigid deformations, low color contrast and transparency.

Direct matching between distant frames can be thus sensitive to ambiguous correspondences. An alternative consists in computing the long-range displacement through concatenation of elementary optical flow fields. This can be done by temporal integration, similarly to dense point tracking algorithms [7]. However, even small errors in the input fields can lead to large drifts in the final motion field.

A first step towards accurate long-range dense correspondences is to combine numerous estimations from direct matching and temporal integration. Following a similar approach to that presented in [10], one can select for each pixel the optimal motion vector among a set of candidate motion fields based on intrinsic motion field quality and spatial regularization. A more sophisticated processing, described in [11, 12], consists in sequentially merging a set of concatenated multi-step motion fields at intermediate frames up to the target frame. However, in either case, the optimal motion vector selection strongly depends on the same optical flow assumptions that frequently fail between distant frames. This issue could be partially compensated by complexifying the matching criteria ad-infinitum, but an uncertainty component is always present. This argues in favor of a statistical processing which takes into account the random nature of these perturbations among a large set of dense motion fields.

In this direction, we propose two main contributions to address the dense matching problem between distant frames. Firstly, we present a combinatorial multi-step integration method which allows one to get a large set of motion fields between two distant frames by considering multiple motion paths across the sequence. Secondly, once this motion candidate construction is performed, we apply a new approach to select the optimal motion field based on statistics and spatial regularization. Results for motion estimation between distant frames in the context of video editing are presented.

2. MOTION CANDIDATE CONSTRUCTION

Let us consider a sequence of \( N + 1 \) RGB images \( \{ I_n \}_{n \in \{0, ..., N\}} \) and let \( I_a \) and \( I_b \) be two distant frames of this sequence with \( 0 \leq a < b \leq N \). In this paper, we focus on the frame pair \( \{ I_a, I_b \} \) and our goal is to accurately estimate a dense motion field between these two frames. In this section, we aim at generating multiple motion maps between \( I_a \) and \( I_b \).

2.1. Direct matching with multiple optical flow estimators

A first approach for building multiple motion candidates consists in considering a direct motion estimation using different optical flow methods [10]. Even if the considered estimators may fail in some regions, the idea is to pool the strengths of each one. Furthermore, the same estimator can be used several times by modifying its parameter settings. In addition,
we can derive from each version a parametric motion field for
which motion is constrained by a global transformation [13].
Direct matching is more ambiguous as the distance between
$I_a$ and $I_b$ increases. Due to a large motion range, the
motion of periodic color patterns or uniform areas may not be
correctly estimated. This supports a motion field construction
stage using concatenation of various optical flow fields.

2.2. Combinatorial multi-step integration

Let us describe the concept of motion path as an alternative
to direct matching for obtaining a displacement map between
$I_a$ and $I_b$. A motion path is obtained through concatenation
of elementary optical flow fields across the video
sequence. It links each pixel $x_a$ of $I_a$ to a corresponding
position in $I_b$. Elementary optical flow fields can be
computed between consecutive frames or with different frame
steps [11, 12], i.e., with 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 $n$. This means that the set of optical
flow fields $\{v_{n,n+s}, v_{n,n+s+2}, \ldots, v_{n,n+s_{Q_n}}\}$ is available.

Our objective is to produce a large set of motion maps
between $I_a$ and $I_b$ as to form a significative set of samples
upon which a statistical processing would be meaningful and
advantageous. Given this objective, we propose to initially
generate all the possible step sequences (i.e., combinations of
steps) in order to join $I_b$ from $I_a$. Each step sequence defines a
motion path. Let $\Gamma_{a,b} = \{\gamma_0, \gamma_1, \ldots, \gamma_{K-1}\}$ be the set of $K$
possible step sequences $\gamma_i$ between $I_a$ and $I_b$. $\Gamma_{a,b}$ is computed
by building a tree structure (Fig. 1) where each node
corresponds to a motion field assigned to a given frame for
a given step value (node value). In pratice, the construction
of the tree is done recursively starting from $I_a$: we create for
each node as many children as the number of steps available
at the current instant. A child node is not generated when $I_b$
has already been reached (therefore, the current node is
considered as a leaf) or if $I_b$ is passed given the considered step.
Finally, once the tree has been built, going from the root
node to leaf nodes gives $\Gamma_{a,b}$, the set of step sequences. For illustration,
the tree in Fig.1 indicates the four step sequences that
can be generated going from $I_0$ to $I_3$ with steps 1, 2 and 3:
$\Gamma_{0,3} = \{\{1,1,1\}, \{1,2\}, \{2,1\}, \{3\}\}$.

Once all the possible step sequences $\gamma_i \in \{0, \ldots, K - 1\}$ between $I_a$ and $I_b$ have been generated, the corresponding
motion paths can be constructed through 1st-order Euler integration.
Starting from each pixel $x_{a}$ of $I_a$ and for each step
sequence, this integration performs the accumulation of optical
flow fields following the steps which form the current step sequence.
Thus, with steps 1, 2 and 3, Fig. 2 illustrates the construction of the four possible motion paths (one for each step sequence of $\Gamma_{0,3}$) between $I_0$ and $I_3$. Let $f_j^i = a + \sum_{k=0}^{j} s_k$ be the current frame number during the construction of motion path $i$ from $I_b$, where $j$ is the step index within the step sequence $\gamma_i$. For each $\gamma_i \in \Gamma_{a,b}$ and for each step $s_j^i \in \gamma_i$, we start from $x_{a}$ in order to iteratively compute:

$$x_{f_j^i} = x_{f_{j-1}^i} + v_{f_{j-1}^i} x_{f_{j-1}^i}$$

(1)

Once all the steps $s_j^i \in \gamma_i$ have been run through, we obtain $x_{f_j^i}$, the corresponding position in $I_b$ of $x_a$ of $I_a$ obtained
with step sequence $\gamma_i$. By considering all the step sequences, we finally get a large set of candidate positions in $I_b$ and this
for each pixel $x_a$ of $I_a$. Note that the occlusion maps attached
to input motion fields are used to possibly stop the motion
path construction. Considering an intermediate point $x_{f_j^i}$ during
the construction, a step can be added only if the closest
pixel to $x_{f_j^i}$ is considered as un-occluded for this step. Otherwise,
the motion path is removed. In the following, the large
set of candidate positions in $I_b$ is defined as $T_{a,b}(x_a) = \{x_{f_j^i}\}$
for all $i \in \{0, \ldots, K_a - 1\}$ where $K_a$ is the cardinal of $T_{a,b}(x_a)$.

Up to now, we have considered an exhaustive generation
of step sequences for clarity. However, very distant frames
and for a large set of steps, it is not possible to consider all
possible step sequences (computational and memory issues).
For instance, for a distance of 30 frames and with steps 1, 2,
5 and 10, the number of possible motion paths is 5877241.
Therefore, the procedure described above is performed on a
reasonable number of step sequences and not for all as previously assumed. Firstly, we limit the number of elementary
vectors composing the motion paths by providing a maximum
number of concatenations $N_s$. Indeed, the concatenation
of numerous vectors may lead to an important drift. Secondly,
we randomly select $N_s$ motion paths among the remaining
motion paths ($N_s$ determined by storage capacity). This selection is guided by the fact that the candidate vectors should
not be highly correlated. The frequency of appearance of a
given step at a given frame must be uniform among all the
possible steps arising from this frame in order to avoid a sys-
tematic bias towards the more populated branches of the tree.
by outliers. Consequently, a robust estimation of the distribution central value is necessary: $Q_{\text{Inc}}$. Thus, $T$ is reversed and started from the previously identified nearest pixel which gives a new candidate for $T_{a,b}(x_a)$. Candidates of $T_{a,b}(x_a)$ which have been obtained through this procedure are defined as reverse. Otherwise, we call them direct.

### 3. MOTION VECTOR SELECTION ON LARGE SETS

#### 3.1. Statistical processing for motion vector selection

Given $T_{a,b}(x_a) = \{x^b_{i}\}_{i \in [0,\ldots,K_x-1]}$, the set of candidate positions in $I_b$ obtained for each pixel $x_a$ of $I_a$ from the motion field construction stage described above, the objective is to select the optimal candidate position $x^b_{\ast} \in T_{a,b}(x_a)$ by exploiting the statistical information on the point distribution and the quality of each candidate. The idea is to assume a Gaussian model for the distribution of $T_{a,b}(x_a)$ and try to find its central value, $x^b_{\ast}$. Using the maximum likelihood estimator (MLE) and imposing the selection among elements of $T_{a,b}(x_a)$, the choice of the optimal candidate position $x^b_{\ast}$ is defined by:

$$x^b_{\ast} = \arg \min_{x^b_{i}} \sum_{j=0}^{K_x-a-1} \left\| x^b_{i} - x^b_{j} \right\|_2^2$$

The assumption of Gaussianity can be largely perturbed by outliers. Consequently, a robust estimation of the distribution central value is necessary:

$$x^b_{\ast} = \arg \min_{x^b_{i}} \med_{j \neq i} \left\| x^b_{i} - x^b_{j} \right\|_2^2$$

Finally, each candidate position $x^b_{i}$ receives a corresponding quality score $Q(x^b_{i})$ computed using the inconsistency values $\text{Inc}(x^b_{i})$. $\text{Inc}(x^b_{i})$ corresponds to the Euclidean distance to the nearest reverse (resp. direct) candidate among the distribution if $x^b_{i}$ is direct (resp. reverse). We aim at assigning high quality to candidates for which the corresponding motion field between $I_a$ and $I_b$ is consistent with a motion field between $I_b$ and $I_a$. Quality scores $Q(x^b_{i})$ are computed as follows: the maximum and minimum values of $\text{Inc}(x^b_{i})$ among all candidates are mapped from 0 to a predefined integer $Q_{\text{max}}$. Intermediate inconsistency values are mapped to the line defined by these two values and the result is rounded to the nearest integer $Q(x^b_{i}) \in [0,\ldots,Q_{\text{max}}]$. The higher $Q(x^b_{i})$, the smaller $\text{Inc}(x^b_{i})$. We aim at promoting candidates in the neighborhood of high quality candidates. In practice, $Q(x^b_{i})$ is used as a voting mechanism [16]: while computing the medians in Equation (3), each sample $x^b_{i}$ is considered $Q(x^b_{i})$ times to set the occurrence of elements $\left\| x^b_{i} - x^b_{j} \right\|_2$ which enforces the forward-backward motion consistency.

The statistical processing being applied for each pixel independently, we describe in what follows a global optimization method which includes regularization.

#### 3.2. Global optimization for motion vector selection

We perform a global optimization stage that fuses for each pixel motion candidates into a single optimal motion field, following the approach of [10]. We introduce $L = \{I_a\}$ as a labeling of pixels $x_a$ of $I_a$ where each label indicates one of the candidates of $T_{a,b}(x_a)$. Let $d_{a,b}^k$ be the corresponding motion vectors of candidates of $T_{a,b}(x_a)$. We define the following energy and minimize it with fusion moves [10, 15].

$$E_{a,b}(L) = \sum_{x_a} \rho_d(C(x_a, d_{a,b}^k(x_a)) + \text{Inc}(x_a + d_{a,b}^k(x_a)))$$

$$+ \sum_{<x_a,y_a>} \rho_r \left( \left\| d_{a,b}^k(x_a) - d_{a,b}^k(y_a) \right\|_2 \right)$$

The data term involves the matching cost $C(x_a, d_{a,b}^k)$ and the inconsistency value $\text{Inc}(x_a + d_{a,b}^k)$ which is introduced to make it more robust. The regularization term involves motion similarities with neighboring positions. $\rho_d$ accounts for local color similarities in frame $I_a$. Functions $\rho_d$ and $\rho_r$
spared three selection procedures: 1) the statistical processing combined to Table 2; 2) the statistical processing combined to global optimization (GO); and 3) the statistical processing combined with the global optimization (SP+GO). Firstly, the final fields of each method have been compared through registration and PSNR assessment. For a given pair \( \{I_a, I_b\} \), the final fields are used to reconstruct \( I_a \) from \( I_b \) through motion compensation and color PSNR scores are computed between \( I_a \) and the registered frame for non-occluded pixels. Tables 1 and 2 show quantitative comparisons through PSNR computed for various distances between \( I_a \) and \( I_b \) respectively on the kiosk of MPI S1 and on whole images of Newspaper. An example of registration of the kiosk for a distance of \( 18 \) frames is provided Fig. 3. Results show that \( SP \) is better than \( GO \) for all pairs. The low diversity of candidates at the output of \( SP \) limits the effect of regularization and explains the slight improvement between \( SP \) and \( SP+GO \). The example of Fig. 3 is interesting due to the temporary occlusion of the kiosk which is jumped by multi-step motion paths. For this complex situation, \( SP+GO \) is more adapted than \( GO \). Secondly, in the context of video editing, we evaluate the accuracy of \( SP+GO \) and \( GO \) by motion compensating in \( I_b \) logos manually inserted in \( I_a \). Fig. 4 presents results for Hope and Newspaper with a distance of 18 and 70 frames respectively. For both cases, \( SP+GO \) shows a clear improvement compared to \( GO \).

The proposed combinatorial integration combined to \( SP+GO \) gives better performance compared to the multi-step fusion (MSF) method [12] according to PSNR scores of Table 2. The MSF method itself has been shown in [12] to outperform state-of-the-art methods such as [3, 6, 7].

### 5. CONCLUSION

We perform long-term dense matching by considering multiple multi-step motion paths along the sequence. Given the resulting large set of motion candidates, we apply a selection procedure where the global optimization stage is preceded by a new statistical processing which exploits the spatial distribution and the intrinsic quality of candidates. It leads to better results compared to state-of-the-art methods.

### Table 2: Registration and PSNR assessment with: the combinatorial integration followed by the global optimization (GO); by the statistical processing combined to GO (SP+GO); the multi-step fusion (MSF) method [12]. PSNR scores are computed on the whole images of Newspaper (Fig. 4).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>21.11</td>
<td>19.33</td>
<td>18.11</td>
<td>17.06</td>
<td>16.29</td>
</tr>
<tr>
<td>SP+GO</td>
<td>21.42</td>
<td>19.53</td>
<td>18.3</td>
<td>17.74</td>
<td>17.09</td>
</tr>
<tr>
<td>MSF [12]</td>
<td>20.5</td>
<td>18.22</td>
<td>17.8</td>
<td>16.95</td>
<td>16.6</td>
</tr>
</tbody>
</table>

### 3.3. Motion vector selection framework

We propose to combine statistical processing and the above global optimization stage to combine simultaneously information about the point distribution, a robust selection based on the intrinsic motion field quality and a spatial regularization. For each \( x_a \in I_a \), the statistical processing is applied to the whole set \( T_a(x_a) \). Then, we select the \( N_{opt} \) best candidates of the distribution with the criterion of median minimization of (3). Finally, fusion moves algorithm fuses by pairs these \( N_{opt} \) candidates up to obtaining the best one.

### 4. RESULTS

Our experiments focus on frame pairs taken from three sequences: MPI S1 [14], Hope and Newspaper. For the selected pairs, the combinatorial multi-step integration has been performed taking input elementary flow fields estimated with a 2D version of the disparity estimator of [17]. For all the experiments, the parameters are: \( N_c = 7 \), \( N_p = 100 \), \( Q_{max} = 2 \), \( N_{opt} = 3 \). Steps 1, 2, 3, 4, 5, 15 and 30 have been used.

After this motion field construction stage, we have compared three selection procedures: 1) the statistical processing (SP), 2) the global optimization (GO) method solved by fusion moves [15], 3) the statistical processing combined with the global optimization (SP+GO).
6. REFERENCES


