Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation

Thomas Brox, Jitendra Malik, *Fellow, IEEE*

Abstract—Optical flow estimation is classically marked by the requirement of dense sampling in time. While coarse-to-fine warping schemes have somehow relaxed this constraint, there is an inherent dependency between the scale of structures and the velocity that can be estimated. This particularly renders the estimation of detailed human motion problematic, as small body parts can move very fast. In this paper, we present a way to approach this problem by integrating rich descriptors into the variational optical flow setting. This way we can estimate a dense optical flow field with almost the same high accuracy as known from variational optical flow, while reaching out to new domains of motion analysis where the requirement of dense sampling in time is no longer satisfied.

Index Terms—motion, video, optical flow

1 Introduction

Motion in the form of optical flow is one of the most dominant bottom-up cues in the visual system of humans and other visual species. It is of great importance for grouping and visual learning, the perception of structure, and for self-localization.

The predominant way to estimate dense optical flow in today's computer vision literature is by variational methods as introduced in the seminal work by Horn and Schunck [18], where a local, gradient-based matching of pixel gray values is combined with a global smoothness assumption. Although the original Horn and Schunck model reveals many limitations in practice, many of them have been tackled by subsequent modifications and extensions of the original model. Motion discontinuities and occlusions can be estimated by employing non-quadratic penalizers in the smoothness term and the data term, respectively [14], [7], [26]. Violations of the constant brightness assumption can be considered by using photometric invariant constraints, such as constancy of the gradient [11], higher order derivatives [28], or color models with photometric invariant channels [27], [37]. Finally, when not linearizing the constancy constraints, the model can deal with large displacements [1]. Proper numerical implementation of such non-linearized models in combination with a continuation method leads to coarse-to-fine warping schemes [26], [11], which have been used much earlier in local techniques, e.g., in the Lucas-Kanade approach [25].

When we say that the *model* of the mentioned approaches can deal with large displacements, we do not say that the final *solution* obtained by these methods reflects this ability in all cases. A decisive problem of variational optical flow, and all methods that introduce global smoothness, is the approximative, local optimization. In the variational setting, the result is biased towards the initialization, which is usually the zero motion field: from all local minima, the approach selects the one with the smallest motion.

The common coarse-to-fine warping scheme relaxes this problem as initial estimates are computed at coarser resolution levels. The motion of larger structures is used as an initial guess for the overall image motion, which is then successively refined by taking into account the evidence of smaller structures. While warping schemes work well in all cases where the small structures move more or less the same way as larger scale structures, the approach is doomed to fail as soon as the relative motion of a small scale structure is larger than its own scale, as shown in Fig. 1. In such a case, the large scale structures predict a motion that is substantially different from the correct one. At the resolution level where the smaller structures appear, local minima prevent the right correction. In coarse-to-fine schemes, the result is no longer biased by the zero motion field but by the motion of the large scale structures.

Situations where the coarse-to-fine heuristic does not work appear quite frequently in practice. Articulated motion in general and human motion in particular are problematic. Small body parts like hands can move extremely fast, hence violating the requirement that the motion of the next larger scale structure is a good indicator for the motion. Many action recognition methods, apart from static cues, rely on optical flow. Clearly, they cannot fully exploit the motion cue, since the optical flow estimated with current methods is unreliable just in situations where it is most informative: when there is a clear, distinct motion of a certain body part.

We consider the failure of contemporary optical flow meth-

- *T. Brox and J. Malik are with the Department of Electrical Engineering and Computer Science, University of California at Berkeley. E-mail: {brox,malik}@eecs.berkeley.edu*

This work was supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD) and by ONR MURI N00014-06-1-0734.

Digital Object Indentifier 10.1109/TPAMI.2010.143 0162-8828/10/$26.00 © 2010 IEEE

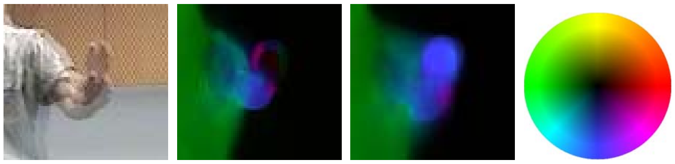

Fig. 1. **Left:** The fast motion of a hand is a typical example where conventional warping methods fail. **Center Left:** Optical flow field computed with such a warping method [11]: the hand motion is missed. **Center Right:** For comparison the optical flow field with the technique presented in this paper: the motion of the hand is estimated correctly. **Right:** Color code for visualizing the flow fields.

ods to reliably capture large displacements as the most limiting factor when applying optical flow in other computer vision tasks. The contribution of the present paper is a variational model and a corresponding numerical scheme that can deal far more reliably with large displacements than previous methods.

The basic idea is to support the continuation method, which is responsible for estimating large displacements in classic warping methods, by another technique that is well known for its ability to estimate arbitrarily large displacements: descriptor matching. In contrast to single pixels, rich local descriptors, such as SIFT or HOG, are usually unique enough to allow for global matching without additional regularity constraints. This renders matching without limitations on the magnitude of the displacement extremely simple and efficient, and explains the enormous success of descriptive features in structure from motion, image search, and object detection.

In optical flow estimation, descriptor matching has not been a success story so far. The reasons for this are quite evident: (i) although most descriptors can be uniquely matched between images, some of them are confused or their counterpart in the other image is missing due to occlusions. This causes a certain amount of mismatches that are very disturbing for most optical flow applications. (ii) Descriptor matching is a discrete technique, which only allows for pixel accuracy. This quantization effect prevents distinguishing small motions and causes drift in tracking applications. (iii) The most successful descriptors are all based on spatial histograms. Histograms are not well localized, and thus the precision of the motion estimates, especially at motion discontinuities, is lower than with, e.g., variational techniques.

One would like to benefit both from the ability of descriptor matching to produce a large amount of correct large displacement correspondences *and* from the ability of variational techniques to efficiently produce highly accurate, dense motion fields without outliers. We achieve this by integrating the correspondences from descriptor matching into a variational optical flow model. As we will describe later in more detail, descriptor matching and the continuation method used as an optimization heuristic in warping techniques are mostly complementary in the way how they avoid local minima in the energy. In conjunction with a coarse-to-fine optimization, descriptor matching can guide the solution towards large displacements of small, independently moving structures, while the other constraints in the variational model successively remove the mismatches and provide the accuracy known from variational methods. Fig. 2 demonstrates that straightforward postsmoothing of descriptor matches or simple initialization of a variational optical flow technique with the descriptor matching result generally does not work. In contrast, the results we obtain with the proposed large displacement optical flow approach prove to be very reliable on a wide variety of video data.

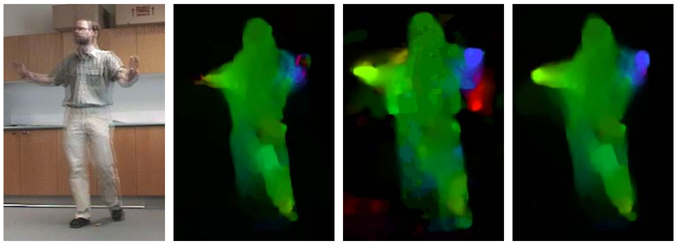

Fig. 2. Straightforward combinations of descriptor matching and variational methods do not work as well as the proposed large displacement optical flow. **Left:** Transparent overlay of input frames. **Center Left:** Initialization of [11] with descriptor correspondences. The initialization is already smoothed away at the coarsest resolution and does not help to estimate the fast hand motion. **Center Right:** Postsmoothing of dense HOG correspondences with TV regularization. Smoothing alone cannot remove all mismatches. Moreover, motion discontinuities are severely dislocated. **Right:** Proposed large displacement optical flow (LDOF).

2 RELATED WORK

The use of richer descriptors in optical flow estimation goes back to Weber and Malik, who employed a multi-scale set of filter responses, so-called jets, in a Lucas-Kanade like setting [33]. The linearization involved in this method keeps it from estimating large displacements. In contrast, Liu et al. [23] have recently proposed a method that computes dense correspondence fields between two different scenes. Clearly, the matching of scenes induces very large displacements and requires invariance to intra-category variations. The idea in [23] is to compute a dense field of SIFT descriptors and then run an approximative discrete optimization via belief propagation from [29] on top of these descriptors. In contrast to simple nearest neighbor matching, SIFT flow tries to minimize an energy that also includes regularity constraints. The model and numerical scheme we present in the present paper differs from SIFT flow in three ways. First, as we focus more on classic motion analysis rather than scene matching, our model does not fully rely on histogram based features such as SIFT. Such features are only a supplement in our approach that allows avoiding local minima, but we still match features such as the color and gradient of single pixels, which have a high spatial resolution. Second, the optimization strategies are different. While SIFT flow considers all possible matches at each pixel when introducing the regularity constraint, our scheme only considers the best matches and checks their consistency with the regularity constraint and the other image features. This makes our optimization much more efficient. Additionally, we need to perform a nearest neighbor search only for a subsampled set of pixels since precise localization is provided by the pixel color and gradient in the variational setting. Third, we have a continuous rather than a discrete model, which provides subpixel accuracy and does not suffer from discretization artifacts, such as motion boundaries being aligned with the coordinate axes.

Another related work is [4], which uses integer quadratic programming (IQP), a graph matching strategy that combines descriptor matching with a regularity constraint. Since even approximations of IQP are computationally demanding, only sparse feature sets can be efficiently matched and interpolated by spline functions. In contrast, the goal of the present paper is to produce a dense flow field including motion boundaries.

Finally, landmarks have been used for a long time in classical registration problems. The important difference in these approaches is that the landmarks are usually manually defined and only suffer from Gaussian noise, whereas the point correspondences in our approach are derived automatically and can be severe outliers. Few exceptions in registration where correspondences are automatically computed can be found in [13], [35]. Another difference is that registration assumes a smooth transformation without motion discontinuities and occlusions. This allows to reduce the number of model parameters considerably, e.g., by using the Thin-Plate-Spline model.

Very related is also the work in [17]. Large displacement motion is estimated based on a correlation term and integrated into a variational model similar to the present work. Main conceptual differences of this work are its focus on atmospheric data and its strict separation of estimating a large displacement motion field and a small displacement increment in a two-stage process.

The presented approach can also be regarded more generally as a combination of discrete optimization in the form of descriptor matching and continuous optimization. Combinations of both forms of optimization have been used, e.g., in [21], [20], where the discrete optimization does the main work, and a continuous postprocessing step finally provides subpixel accuracy. In contrast, the present approach couples both optimization strategies more closely.

The present paper extends a preliminary conference version [10], where region correspondences serve to recover large displacements. We investigate other features to provide point correspondences and compare their suitability in the optical flow setting. We also modified details that lead to a better overall performance. Moreover, we tested the method on a much larger variety of videos.

3 Variational model

Let $\mathbf{I}_1, \mathbf{I}_2 : (\Omega \subset \mathbb{R}^2) \to \mathbb{R}^d$ be the first and the second frame to be aligned. For a gray scale image we have $d = 1$ and for color images $d = 3$. Moreover, $\mathbf{x} := (x, y)^\top$ denotes a point in the image domain Ω, and $\mathbf{w} := (u, v)^\top$ is the optical flow field, i.e., a function $\mathbf{w} : \Omega \to \mathbb{R}^2$. A common assumption is that corresponding points should have the same gray value or color. This can be expressed by the energy

$$E_{\text{color}}(\mathbf{w}) = \int_\Omega \Psi\left(|\mathbf{I}_2(\mathbf{x} + \mathbf{w}(\mathbf{x})) - \mathbf{I}_1(\mathbf{x})|^2\right) d\mathbf{x} \quad (1)$$

which penalizes deviations from this assumption. Note that in contrast to the Horn and Schunck model, there is no linearization involved here, which enables the estimation of large displacements. The robust function $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}, \epsilon = 0.001$ allows to deal with occlusions and other non-Gaussian deviations of the matching criterion. It corresponds to a Laplace distribution which has longer tails than the Gaussian distribution. In other works, even longer-tailed distributions have been advocated [6], [7]. The advantage of the Laplace distribution is that the corresponding penalizer is still convex, simplifying the optimization.

Due to illumination effects, matching the color or gray value is not always reliable. Therefore, it has been suggested to supplement the constraint in (1) by a constraint on the gradient, which is invariant to additive brightness changes [11]:

$$E_{\text{grad}}(\mathbf{w}) = \int_\Omega \Psi\left(|\nabla\mathbf{I}_2(\mathbf{x} + \mathbf{w}(\mathbf{x})) - \nabla\mathbf{I}_1(\mathbf{x})|^2\right) d\mathbf{x}. \quad (2)$$

Other higher order constraints have been investigated in [28] showing that the gradient constraint works best, as it introduces the required invariance properties without being as sensitive to noise as second order constraints. An alternative feature with similar effects can be obtained by structure-texture decomposition [34].

Both (1) and (2) enforce the matching of only weakly descriptive features. Just optimizing the sum of these two energies would result in many ambiguous solutions, most of them not being consistent with the true optical flow. This underlines the power and importance of regularity constraints in optical flow estimation. Regularity can be enforced, for instance, by penalizing the total variation of the flow field:

$$E_{\text{smooth}}(\mathbf{w}) = \int_\Omega \Psi\left(|\nabla u(\mathbf{x})|^2 + |\nabla v(\mathbf{x})|^2\right) d\mathbf{x}. \quad (3)$$

Putting all these constraints together yields the model presented in [11], [12]:

$$E(\mathbf{w}) = E_{\text{color}} + \gamma E_{\text{gradient}} + \alpha E_{\text{smooth}}. \quad (4)$$

From the modelling point of view, this model is extremely general. It can deal with all kinds of deformations, motion discontinuities, occlusions, and arbitrarily large displacements. The reason why optical flow estimation is not a perfectly solved problem yet is due to the approximative optimization of this energy model, not the model itself.

The optimization strategy in variational optical flow is all based on local optimization in conjunction with a coarse-to-fine method. Initial estimates are obtained by removing detail from the input data. Such a strategy puts much emphasis on large structures. In this paper, the idea is to supplement a complementary approximation to the optimization scheme, namely to remove global regularity. The color and gradient at a point is not a descriptive feature. It becomes descriptive due to the global regularity constraint, but at the cost of generating a hard optimization problem. Alternatively, we can use more descriptive features and neglect the regularity constraint. Point matching without a regularity constraint can be done efficiently in a globally optimal manner by simple nearest neighbor matching.

In the end, however, we would like to enforce a smooth flow field. Moreover, descriptor matching has important drawbacks. First, it is a discrete method that does not provide subpixel accuracy. Second, the fixed spatial extent of rich descriptors is responsible for inaccuracies at motion discontinuities and in case of all non-translational motions. For these reasons, we would like to combine descriptor matching with the variational model and its coarse-to-fine optimization. To this end, we integrate the point correspondences from descriptor matching into the variational approach by adding another term:

$$E_{\text{match}}(\mathbf{w}) = \int \delta(\mathbf{x})\rho(\mathbf{x})\Psi\left(|\mathbf{w}(\mathbf{x}) - \mathbf{w}_1(\mathbf{x})|^2\right)\, d\mathbf{x}. \quad (5)$$

In this term, $\mathbf{w}_1(\mathbf{x})$ denotes the correspondence vectors obtained by descriptor matching at some points $\mathbf{x}$. $\delta_i(\mathbf{x})$ is 1 if there is a descriptor available in frame 1 at point $\mathbf{x}$; otherwise it is 0. Each correspondence is weighted by its matching score $\rho_i(\mathbf{x})$, which will be exactly defined in section 4.1. The same robust function Ψ as above is applied in order to allow for wrong matches. As suggested in [10], one could also integrate the best K matches. However, it turns out that increasing the number of hypotheses has actually more drawbacks than advantages as explained in detail in the appendix.

Eq. (5) assumes that the descriptors are already matched. We can formulate this matching task as another energy term to be minimized:

$$E_{\text{desc}}(\mathbf{w}_1) = \int \delta(\mathbf{x})\, |\mathbf{f}_2(\mathbf{x} + \mathbf{w}_1(\mathbf{x})) - \mathbf{f}_1(\mathbf{x})|^2\, d\mathbf{x}, \quad (6)$$

where $\mathbf{f}_1(\mathbf{x})$ and $\mathbf{f}_2(\mathbf{x})$ denote the (sparse) fields of feature vectors in frame 1 and frame 2, respectively. Plugging all terms together, we can state the whole model as a single optimization problem:

$$\begin{aligned} E(\mathbf{w}) = E_{\text{color}}(\mathbf{w}) + \gamma E_{\text{gradient}}(\mathbf{w}) + \alpha E_{\text{smooth}}(\mathbf{w}) \\ + \beta E_{\text{match}}(\mathbf{w}, \mathbf{w}_1) + E_{\text{desc}}(\mathbf{w}_1), \end{aligned} \quad (7)$$

where α, β, and γ are tuning parameters which can be determined manually according to qualitative evidence on a large variety of videos, or be estimated automatically from ground truth data [32].

On the first glance, this energy looks more complex than necessary. (a) Why adding two terms E_{match} and E_{desc} with the auxiliary variable $\mathbf{w}_1$ rather than directly adding $E_{\text{desc}}(\mathbf{w})$? (b) Why still using the color and gradient features, when there are more descriptive features involved?

Concerning (a), the auxiliary variable allows to integrate discrete descriptor matching into a continuous approach in the form of soft constraints; see also [31]. Without this auxiliary variable and the coupling term E_{match}, discrete matching would not be compatible with the variational setting.

Concerning (b), rich descriptors have drawbacks when spatial localization is concerned. Hence, point features like color and gradient are actually preferable. In fact, the effect of the additional terms E_{match} and E_{desc} disappears in the continuous limit. This is easy to see. The descriptors are only available on a fixed spatial grid defined by the δ function. Aside the grid points, these terms are zero. The other features and the smoothness term on the other hand are defined in the continuous domain with the number of point features going to infinity. In the coarse-to-fine minimization process, the additional terms only affect the energy at coarser levels. The final continuous energy is unaffected. The minimization framework will now be explained in detail.

4 Minimization

The final goal is to find a minimum equal or similar to the global minimum of the energy in (7), which in the continuous limit is equivalent to the energy in (4). Since this functional is highly non-convex, we need reasonable approximation schemes that find a good initial guess of the solution.

We rely here on a combination of two methods that produce initial guesses: (a) descriptor matching and (b) a continuation method in the spirit of graduated non-convexity [8]. Both methods approximate the energy by simpler versions of the energy which can be globally optimized, yet they are complementary in the way how they simplify the energy. Descriptor matching neglects regularity, whereas the continuation method neglects image details. The results from descriptor matching can be integrated in the continuation method, thus we start with descriptor matching and explain the continuation method afterwards.

4.1 Descriptor matching

The descriptor matching part focuses on minimizing $E_{\text{desc}}(\mathbf{w}_1)$ independently from the rest of the energy. Decoupling $E_{\text{desc}}(\mathbf{w}_1)$ enables global optimization of this subproblem.

Proposition 1: *For descriptors given on discrete grids in frame 1 and frame 2, minimizing $E_{\text{desc}}(\mathbf{w}_1)$ with respect to $\mathbf{w}_1$ is a discrete optimization problem. Global optimization can be achieved by complete search with complexity $\mathcal{O}(mn)$, where m and n denote the cardinalities of the grids in frame 1 and frame 2, respectively.*

Proof: Let $\delta(\mathbf{x})$ define a discrete grid in frame 1 and $\delta'(\mathbf{x})$ another grid in frame 2. Usually, δ' will be a finer grid than δ. We can rewrite

$$\begin{aligned} E_{\text{desc}}(\mathbf{w}_1) &= \int \delta(\mathbf{x})\, |\mathbf{f}_2(\mathbf{x} + \mathbf{w}_1(\mathbf{x})) - \mathbf{f}_1(\mathbf{x})|^2\, d\mathbf{x} \\ &= \sum_{i,\delta(\mathbf{x}_i)=1} |\mathbf{f}_2(\mathbf{x}_i + \mathbf{w}_1(\mathbf{x}_i)) - \mathbf{f}_1(\mathbf{x}_i)|^2 \end{aligned} \quad (8)$$

Clearly, due to a missing regularity constraint, the sum entries are mutually independent. Thus, we can optimize $\mathbf{w}_1$ at each grid point $\mathbf{x}_i$ independently. This can be achieved by evaluating the energy for all possible grid points x_j of δ' and choosing x_j for which this energy is minimal. The optimal $\mathbf{w}_1(\mathbf{x}_i) = \mathbf{x}_j - \mathbf{x}_i$. Obviously, the overall time complexity is $\mathcal{O}(mn)$. It can be reduced by using efficient nearest neighbor search [16]. ■

With the optimization of this part being extremely simple, we only need to define reasonable descriptors and grids where these descriptors are available. We investigate here three different methods: one based on region matching as proposed in [10], one based on HOG descriptors [15], and one based on geometric blur (GB) [5]. The main requirements for a good descriptor matching method is that the grid is fine enough to capture the motion of smaller structures, and that the descriptors are unique enough to limit the number of false matches.

4.1.1 Region matching

For creating regions in the image, we rely on the segmentation method proposed in [2]. It creates a hierarchical oversegmentation of the image. For each region of this hierarchy, we can compute a descriptor based on SIFT and color and use these descriptors for matching. For details about building the descriptors from the regions and the exact matching procedure we refer to [10]. The centroids of the region i in the first frame and the matched region j in the second frame serve as the grid points x_i and x_j, where the region correspondence is integrated into the variational energy. The matching score is defined as:

$$\rho(\mathbf{x}_i) := \frac{d_2 - d_1}{d_1}, \tag{9}$$

where d_1 and d_2 denote the distances of the best and the second best match, respectively. The distances are the sums of squared differences of warped patches; see [10] for details. Taking the second best match into account, helps to give more weight to unique matches, whereas unclear situations get damped.

For the experiments in Section 5, we supplement a consistency check to the original matching procedure in [10]. A region correspondence will be counted only if the best match from frame 1 to frame 2 is equal to the best match from frame 2 to frame 1.

4.1.2 Histogram of oriented gradients

As an alternative to building descriptors from regions, we densely compute histograms of oriented gradients (HOG) in both frames [15]. Each gradient histogram comprises 15 different orientations and is computed in a 7×7 neighborhood. In contrast to [15], the sign of the gradient is not neglected. The computation is very fast when using integral images. We also apply a Gaussian filter with $\sigma = 0.8$ in orientation direction to reduce quantization effects.

The final descriptors are computed by collecting the histograms at the central pixel and the eight neighbors in a distance of 4 pixels; cf. Fig. 3. Thus, each descriptor consists of $15 \cdot 9 = 135$ entries, and descriptors are available at every pixel position in frame 1 and frame 2. The distance is computed as the sum of squared differences between the two vectors, and the matching score ρ is defined the same way as in (9).

We define the grid $\delta(\mathbf{x})$ by picking a descriptor at every fourth pixel in x- and y-direction. This reduces the matching effort by a factor 16 compared to sampling a descriptor at every pixel. Since every histogram has a 7×7 support, this subsampling does not miss any small structures. Moreover, it is worth noting that the full resolution needed for precise localization of motion boundaries is provided by the color and gradient cues in the variational setting. These are better suited to ensure high precision solutions, as HOG descriptors are unprecise anyway due to their histogram nature.

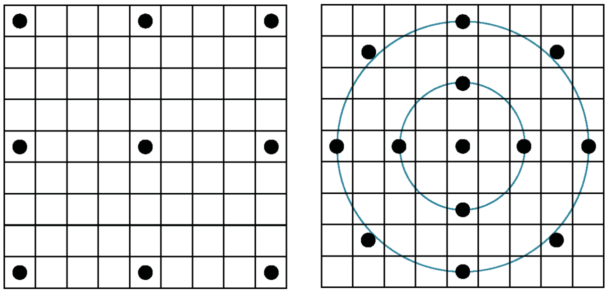

Fig. 3. **Left:** Each HOG descriptor consists of 9 local histograms sampled at equidistant points. **Right:** Each GB descriptor consists of local histograms computed at three different integration scales.

We also compute the smaller eigenvalue λ of the structure tensor $\nabla I \nabla I^\top$ integrated over the same area as the histograms, and ignore descriptors at points where λ is smaller than one eighth of the average across the whole image. The incentive is to save computation time and to reduce the number of false matches by ignoring areas without any structure.

The grid $\delta'(\mathbf{x})$ is defined at the full pixel resolution to ensure pixel accurate matches. For each match $(\mathbf{x}_i, \mathbf{x}_j)$, we check whether $(\mathbf{x}_j, \mathbf{x}_i)$ is also the best match in backward direction. If not, $\delta(\mathbf{x}_i)$ is set to 0. This consistency check removes many false matches, particularly those due to occlusion.

It is worth noting that the HOG descriptor is very similar to the SIFT descriptor from [24] as soon as the scale and rotation invariance of the SIFT detector is neglected. Both descriptors can be made equivalent by sampling 8 orientations on a 4 by 4 grid and using a Gaussian kernel rather than a box kernel. In the optical flow setting, we found it advantageous to have a slightly reduced spatial extent of the descriptor, as this reduces the blurring effects at motion discontinuities.

4.1.3 Geometric blur

Any other rich descriptor is applicable as long as it can be computed densely across the image. As an alternative to HOG/SIFT-like descriptors, we tried a variant of geometric blur (GB) from [5]. We compute 15 oriented gradients like in the HOG descriptor above, but rather than building a histogram by applying a 7×7 box kernel, we apply Gaussian blurring with three different blur levels $\sigma_0 = 0$, $\sigma_1 = 1$, and $\sigma_2 = 2$. The descriptor is assembled of one entry from level 0, 4 entries from level 1, and 8 entries from level 2 as illustrated in Fig. 3. The sampling grids in the two frames and the matching procedure are exactly the same as described above for the HOG descriptors.

4.2 Continuation method

After decoupling $E_{\text{desc}}(\mathbf{w}_1)$, the remainder of (7) can be minimized the same way as proposed in [11] using a continuation method. The additional term E_{match} is convex in $\mathbf{w}$ and does not cause any trouble. The idea of the continuation method is to split the original problem into a sequence of subproblems at different resolution levels by smoothing the input images. In order to allow for smooth transitions between levels, we use a very fine pyramid, where the image at level k is a downsampled version of the input image with downsampling factor $0.95^{(k_{\max}-k)}$. $k_{\max}$ is chosen such that discrete derivative filters can still be applied. Another way of continuation is by just smoothing the input images using a Gaussian kernel but keeping the original resolution [1]. Continuation methods have been used also in many other contexts, e.g. [8].

Proposition 2: *The subproblem in each continuation step is convex and can be globally optimized for fixed correspondences* $\mathbf{w}_1$.

Proof: The Euler-Lagrange equations for (7) read:

$$\begin{aligned}
&\Psi'\left(I_z^2\right) I_z I_x + \gamma\Psi'\left(I_{xz}^2 + I_{yz}^2\right)\left(I_{xx}I_{xz} + I_{xy}I_{yz}\right)\\
&\quad+\beta\,\rho\,\Psi'\left((u-u_1)^2 + (v-v_1)^2\right)(u-u_1)\\
&\quad-\alpha\,\mathrm{div}\left(\Psi'\left(|\nabla u|^2 + |\nabla v|^2\right)\nabla u\right) = 0\\
&\Psi'\left(I_z^2\right) I_z I_y + \gamma\Psi'\left(I_{xz}^2 + I_{yz}^2\right)\left(I_{xy}I_{xz} + I_{yy}I_{yz}\right)\\
&\quad+\beta\,\rho\,\Psi'\left((u-u_1)^2 + (v-v_1)^2\right)(v-v_1)\\
&\quad-\alpha\,\mathrm{div}\left(\Psi'\left(|\nabla u|^2 + |\nabla v|^2\right)\nabla v\right) = 0,
\end{aligned} \tag{10}$$

where $\Psi'(s^2)$ is the first derivative of $\Psi(s^2)$ with respect to s^2, and we define

$$\begin{array}{ll}
I_x := \partial_x I_2(\mathbf{x}+\mathbf{w}) & I_{xy} := \partial_{xy} I_2(\mathbf{x}+\mathbf{w})\\
I_y := \partial_y I_2(\mathbf{x}+\mathbf{w}) & I_{yy} := \partial_{yy} I_2(\mathbf{x}+\mathbf{w})\\
I_z := I_2(\mathbf{x}+\mathbf{w}) - I_1(\mathbf{x}) & I_{xz} := \partial_x I_z\\
I_{xx} := \partial_{xx} I_2(\mathbf{x}+\mathbf{w}) & I_{yz} := \partial_y I_z.
\end{array} \tag{11}$$

Nested fixed point iterations can be used to resolve the nonlinearity in (10). The continuation method is integrated into this numerical scheme by running the outer iteration loop across multiple image scales. The initialization $\mathbf{w}^0 := (0,0)$ is specified at the coarsest scale, and updates $\mathbf{w}^{k+1} = \mathbf{w}^k + \mathbf{dw}^k$ are computed at successively finer scales, where $\mathbf{dw}^k := (du^k, dv^k)$ is the solution of

$$\begin{aligned}
&\Psi_1' I_x^k (I_z^k + I_x^k du^k + I_y^k dv^k) + \beta\rho\Psi_3'(u^k + du^k - u_1)\\
&\quad+\gamma\,\Psi_2' I_{xx}^k (I_{xz}^k + I_{xx}^k du^k + I_{xy}^k dv^k)\\
&\quad+\gamma\,\Psi_2' I_{xy}^k (I_{yz}^k + I_{xy}^k du^k + I_{yy}^k dv^k)\\
&\quad-\alpha\,\mathrm{div}\left(\Psi_4'\nabla(u^k + du^k)\right) = 0\\
&\Psi_1' I_y^k (I_z^k + I_x^k du^k + I_y^k dv^k) + \beta\rho\Psi_3'(v^k + dv^k - v_1)\\
&\quad+\gamma\,\Psi_2' I_{xy}^k (I_{xz}^k + I_{xx}^k du^k + I_{xy}^k dv^k)\\
&\quad+\gamma\,\Psi_2' I_{yy}^k (I_{yz}^k + I_{xy}^k du^k + I_{yy}^k dv^k)\\
&\quad-\alpha\,\mathrm{div}\left(\Psi_4'\nabla(v^k + dv^k)\right) = 0
\end{aligned} \tag{12}$$

with

$$\begin{aligned}
\Psi_1' &:= \Psi'\left((I_z^k + I_x^k du^k + I_y^k dv^k)^2\right)\\
\Psi_2' &:= \Psi'\big((I_{xz}^k + I_{xx}^k du^k + I_{xy}^k dv^k)^2\\
&\qquad + (I_{yz}^k + I_{xy}^k du^k + I_{yy}^k dv^k)^2\big)\\
\Psi_3' &:= \Psi'\left((u^k + du^k - u_1)^2 + (v^k + dv^k - v_1)^2\right)\\
\Psi_4' &:= \Psi'\left(|\nabla(u^k + du^k)|^2 + |\nabla(v^k + dv^k)|^2\right).
\end{aligned} \tag{13}$$

It can be verified that the equations in (12) are the Euler-Lagrange equations of the energy:

$$\begin{aligned}
E^k(du^k, dv^k) = &\int_\Omega \Psi\left((I_x^k du^k + I_y^k dv^k + I_z^k)^2\right)\, d\mathbf{x}\\
&+\gamma\int_\Omega \Psi\left((I_{xx}^k du^k + I_{xy}^k dv^k + I_{xz}^k)^2\right)\, d\mathbf{x}\\
&+\gamma\int_\Omega \Psi\left((I_{xy}^k du^k + I_{yy}^k dv^k + I_{yz}^k)^2\right)\, d\mathbf{x}\\
&+\beta\int_\Omega \Psi\left((u^k + du^k - u_1)^2 + (v^k + dv^k - v_1)^2\right)\, d\mathbf{x}\\
&+\alpha\int_\Omega \Psi\left(|\nabla(u^k + du^k)|^2 + |\nabla(v^k + dv^k)|^2\right).
\end{aligned} \tag{14}$$

With $\Psi(s^2)$ being convex in s, each of the terms is convex in (du^k, dv^k), and consequently the energy E^k is convex for all k and can be globally optimized. ■

For solving (12), an inner fixed point iteration over l is employed, where the robust functions in (13) are set constant for fixed $du^{k,l}$, $dv^{k,l}$ and are iteratively updated. The equations are then linear in $du^{k,l}$, $dv^{k,l}$ and can be solved by standard iterative methods after proper discretization. The fact that we discretize the equation system rather than the energy model allows for a consistent discretization, i.e., one that converges to the continuous limit when the image resolution gets finer. For this reason there are no discrete artifacts like edges being aligned with the grid axes. Discretization details can be found, e.g., in [9].

The fact that the full optimization problem in (7) can be split into a number of subproblems that can all be optimized globally does *not* guarantee a global optimum for the full problem. This is well known also in other optimization settings, for instance expectation-maximization. However, the proposed optimization procedure helps to avoid most of the local minima of the original problem. It is worth noting that this is not just a post-smoothing of the correspondences from descriptor matching, but the correspondences are integrated into the continuation method. This gives them high impact at the beginning of the process, where the image resolution is very small and the correspondences dominate the gray value and gradient constancy terms. As the resolution increases, the ratio between the fixed number of point correspondences and the increasing number of pixels in the image drops, and so does the impact of the correspondences. In the continuous limit, this ratio goes to zero. We simulate this limit by running one last iteration with $\beta = 0$.

The naturally decreasing importance of descriptor matching at finer levels of the continuation method is very helpful in practice. At coarse levels, the high impact of the matches pushes the estimation towards large displacement solutions which are otherwise ignored by the warping scheme. At finer levels, this effect is no longer needed, whereas false matches need to be sorted out at some point. Fig. 4 shows the evolution of the solution in a sample case.

One could imagine iterating the whole optimization process, thereby providing feedback to the descriptor matching by restricting the search range of $\mathbf{w}_1$, for instance. This has been proposed in [31], where a continuation method is not

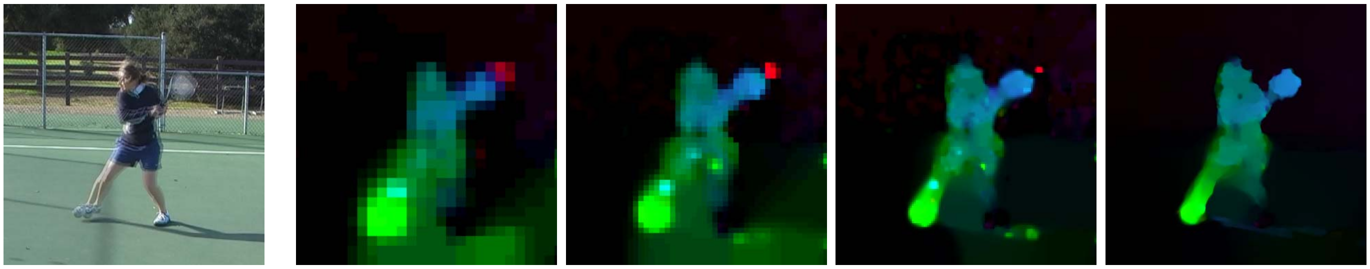

Fig. 4. Evolution of estimated flow. **Left:** Overlayed input images. **Right:** Evolving flow field from coarse (left) to fine (right). The correspondences dominate the estimate at the beginning, pushing the solution towards the fast motion of the leg and the racket. Some wrong matches are also visible, e.g. at the tip of the racket. These outliers are removed over time as more and more data from the image is taken into account.

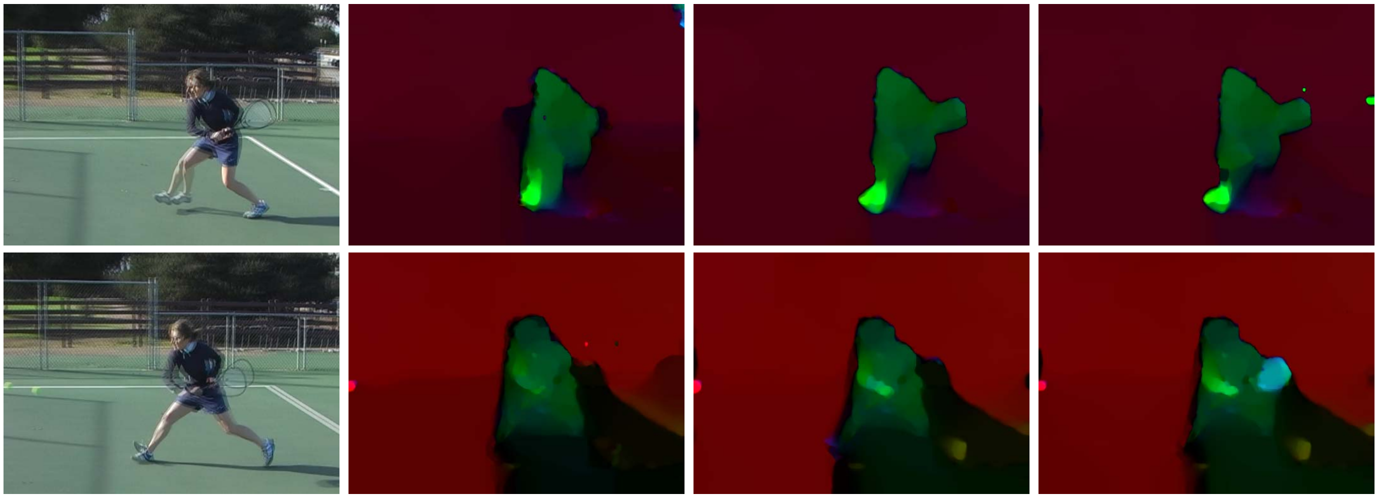

Fig. 5. Comparison of descriptors used for matching. **From left to right:** Overlaid input images, region descriptors as proposed in [10], HOG descriptors, GB descriptors. The region descriptors produce more mismatches than HOG descriptors and miss more parts like the racket than HOG and GB. GB is best in recovering all parts, like the racket in the second example, whereas HOG produces fewer false matches.

used at all. In the setting here, further iterations yield only very little improvement, but increase the computational cost considerably.

5 Experiments

5.1 Comparison among the used descriptors

As we suggested three alternative ways for descriptor matching, in a first experiment we evaluated which one works best. For a quantitative measurement, we ran the methods on all 8 sequences of the Middlebury benchmark with public ground truth [3]. It is important to note that the Middlebury benchmark does not include any ground truth examples with large motion. All the examples can be easily handled with conventional warping techniques. The additional descriptor matching cannot be expected to improve the accuracy in the case of small displacements, as it usually produces some disturbing false large displacement matches, while the correct matches do not have positive effects as the warping already produces very good solutions with subpixel accuracy. Therefore, this experiment cannot tell which of the descriptors is best for dealing with large displacement situations, but which one produces the least false matches. By comparing the numbers to the baseline method without descriptor matching, i.e. $\beta = 0$, we can also measure the accuracy that is lost by adding the ability to deal with large displacement scenarios.

The parameters σ (presmoothing of the images), α, and γ were optimized as to produce the best average angular error among all 8 sequences. $\beta = 300$ was kept at the same value as in all the other examples, ensuring that fast motion could be estimated, if it was present in the sequences. Table 1 shows the average angular error. As expected, the baseline method performs best on this benchmark. Among the descriptor matching techniques, the HOG descriptor leads to the smallest loss in accuracy, followed by GB and region matching. With 16%, the loss in accuracy is a price worth paying for the ability to capture much larger displacements.

The conjecture that HOG descriptors lead to the smallest number of mismatches is also confirmed by a qualitative analysis. Fig. 5 shows two examples from a tennis sequence including large displacements. Both region matching and GB descriptors lead to some artifacts in the final flow that result from false descriptor matching and that could not be pruned by the variational method, whereas the result with HOG matching

	Warping only ($\beta = 0$)	Regions	HOG	GB
Dimetrodon	1.82	1.74	1.85	1.95
Grove2	2.09	2.25	2.68	2.79
Grove3	5.59	6.55	6.38	6.35
Urban2	2.28	3.05	2.64	3.15
Urban3	3.99	5.76	5.07	5.19
RubberWhale	3.77	3.84	3.94	4.14
Hydrangea	2.32	2.36	2.44	2.54
Venus	5.19	7.37	6.45	6.52
Average	**3.38**	**4.11**	**3.93**	**4.08**

TABLE 1
Quantitative comparison of the descriptors on sequences of the Middlebury benchmark (average angular error). On average, HOG descriptors lead to the smallest loss in accuracy.

σ	α	β	γ	AAE
0.6	9	300	3	3.93°
0.3				4.25°
1.2				4.54°
	4.5			4.31°
	18			4.40°
			1.5	4.07°
			6	4.08°

TABLE 2
Effect of parameter variation. Results are quite stable.

	Warping [11]	LDOF	SIFT flow [23]
Fig.6, row 1, 384 × 288	7s	18s	99s
Fig.6, row 2, 373 × 485	12s	29s	167s
Fig.6, row 3, 450 × 350	10s	44s	144s
Fig.6, row 4, 640 × 480	21s	80s	67s*
Fig.6, row 5, 530 × 380	13s	39s	177s

TABLE 3
Computation times on various image sizes. Due to insufficient memory, the fourth entry in the SIFT flow column was obtained on half the resolution. Large displacement optical flow (LDOF) is up to 5 times faster than SIFT flow.

does not show these artifacts. On the other hand, the second example reveals that HOG matching misses the motion of the racket, which is captured correctly with the GB descriptors. Region matching misses the motion of the racket as well as the arm and the foot.

This qualitative behavior persists when analyzing more frames: HOG descriptors produce the fewest mismatches, whereas GB descriptors tend to capture more details. These effects are most likely due to a different behavior of the descriptors in the consistency check, as the integration areas of both descriptors are similar and their weighting in the energy with $\beta = 400$ is the same. Due to the better localization of the GB descriptor close to its central point, a match in one direction is more often the best match in opposite direction. As a consequence, more matches pass the consistency check, leading both to more false matches and more correct large displacement matches. The inferior results of region matching with respect to both aspects is due to the sparser sampling of region descriptors, thus missing more parts, and consistency problems between the segmentations of the two frames, which lead to additional false matches.

In all the other experiments, we used the HOG descriptors for matching. Besides the nice property to be more conservative with respect to false matches, the computation of the HOG descriptors is also the most efficient one. In applications where large displacements clearly dominate, GB descriptors could be advantageous, and there are also possibilities to combine both descriptors. However, we did not further investigate this in detail.

Table 2 shows results of an experiment on parameter variation. The method is fairly robust to small changes in the tuning parameters. Even more important: with a fixed set of parameters, the approach can produce reasonable flow estimates on a variety of sequences. In the remaining experiments, we fixed all parameters at $\sigma = 0.8$, $\alpha = 30$, $\beta = 300$, and $\gamma = 5$. It is worth noting that this set of parameters puts more emphasis on smoothness than the above parameters optimized for the Middlebury data. The reason is that flow on the Middlebury sequences is very easy to compute, with rich structure at all scales. According to our experience, optimum parameters on the Middlebury set do not work well on a larger range of real-world sequences.

5.2 Comparison to classical warping and SIFT flow

We compared the LDOF technique on some large displacement examples to a classical warping technique [11] and to SIFT flow [23]. For the latter comparison we used the code provided on the authors' website, which corresponds to their far more efficient coarse-to-fine version from [22]. We removed the lines in their code that downsample the input images to allow for a fair comparison based on the same image resolution.

Fig. 6 shows the computed flows and Table 3 lists the computation times on one core of a Core Duo 2.5GHz laptop with 3GB of memory running Windows Vista. While the descriptor matching clearly takes some extra time compared to [11], it is faster than SIFT flow.

All examples show large displacements that cause problems to classic warping methods. The fast motion of the hands in the first, the leg in the second, and the balls in the fourth and the last example are missed. The hand motion in the fourth example is underestimated. The fast motion of the car in the third example was coarsely captured, because the car itself is a large structure in the image. However, we had to specifically tune the smoothness parameter α to achieve this result, and the reduced smoothness leads to unpleasant artifacts in the flow field. Interestingly, the hand motion in the tennis example is estimated correctly, since the contrast of the hand is very high. On the other hand, the slower foot motion is missed due to lower contrast. These examples clearly demonstrate the necessity for descriptor matching when dealing with optical flow in practice.

With the proposed LDOF method the large displacements are correctly estimated. The large displacements are also well estimated with SIFT flow except for the motion of the right ball in the fourth and the ball in the last example. Comparing LDOF and SIFT flow indicates a higher precision of LDOF. There are two reasons for that. Firstly, SIFT flow shows typical discretization and quantization artifacts, in particular discontinuities aligned with the grid axes and block effects

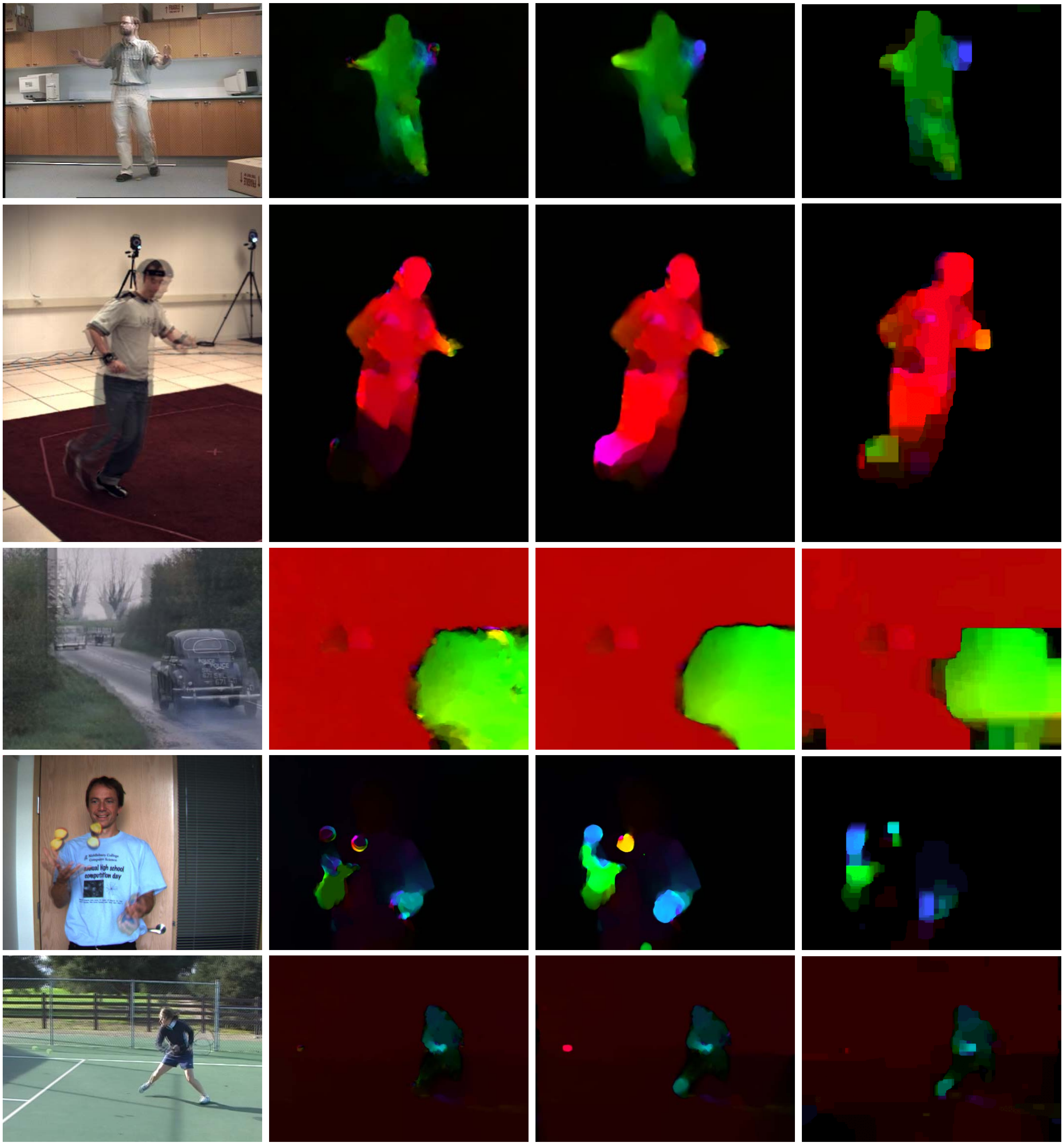

Fig. 6. **Left:** Input images. The second row example is from [30], the last row example from [3]. **Center left:** Classic warping [11]. **Center right:** Proposed large displacement optical flow (LDOF). **Right:** SIFT flow [23].

due to missing subpixel accuracy. Secondly, SIFT and HOG descriptors are histograms and are not perfectly localized. While SIFT flow solely relies on these histogram descriptors, our method uses these descriptors only for pushing the solution towards large displacements, but finally makes use of the well-localized point-wise color and gradient features.

A further advantage of the proposed optimization strategy over belief propagation used in SIFT flow is the efficiency in terms of memory. While our method just requires 120MB of memory to run the VGA example in the last row of Fig. 6, a system with 3GB of memory was not enough to run this example with the coarse-to-fine belief propagation from [22][1].

[1] Fairness requires to mention that part of this problem was due to Windows Vista, which is quite famous for its inefficiency.

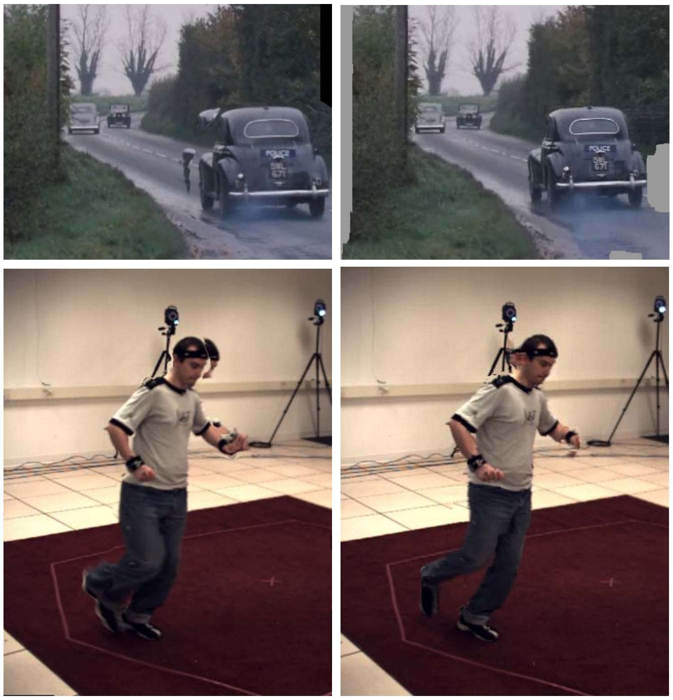

Fig. 7. Warped second image; see Fig. 6 for flow. **Left:** Large displacement optical flow. **Right:** SIFT flow. The SIFT flow warps look nicer as occluding structures are not doubled, but this reveals actual problems in the estimated flow (see text). Animations are available in the supplemental material.

This is why we ran this image pair with half the resolution.

The car example in Fig. 6 demonstrates two limitations inherent to both optimization techniques. The first one is the blurring effect in the lower left corner. Both approximative optimization strategies show similar effects, whereas the global optimum would certainly prefer to match the fine structures on the street correctly[2]. Another problem is the precision of motion boundaries when the background region is weakly textured. In these cases, the motion boundary is dislocated towards the background region. The situation looks better for the SIFT flow result on the first glance because the shape is more car-like, but actually it is worse. This can be seen from Fig. 7. Aside from the fact that the warped car is too small, i.e., SIFT flow underestimates its true motion, the warped image obtained with SIFT flow looks perfectly similar to the first image, pretending a perfect match. However, there is significant occlusion, and a reasonable prior would be that occluded background pixels should move in a similar way as the neighboring background pixels. This would keep parts of the running person and the moving car at their original place and make them visible twice. This is almost achieved by our method, although the space between the double structures still indicates a dislocation of the correct motion boundary. In the presence of occlusions or other mismatches like the tip of the foot in the second example of Fig. 6, SIFT flow is perfect in making the warped second image look like the first one, but the flow at this point is wrong. This behavior is very reasonable for object and scene matching, but it is not useful in the scope of motion analysis. If one is interested in the motion field, it is not recommendable to judge the result only from the warped image. A simple pixel-based nearest neighbor matching can produce the perfect warped image, but the correspondence field is most likely to be completely useless.

5.3 Large displacement optical flow on a number of challenging sequences

To conclude the experiments, we show LDOF on a number of longer sequences taken by ourselves using a consumer camera or grabbed from movies. The input frames and videos of the full sequences with the corresponding optical flow can be downloaded from the first author's home page.

Fig. 8 and 9 show two shots from a Miss Marple movie. Basically in all movies there are some scenes with fast motion that cannot be estimated with conventional techniques. Fig. 8 also shows another comparison to classical warping [11] and SIFT flow [23]. Clearly, at the beginning of the shot, where the motion is small, the warping technique and LDOF yield basically the same results. Versus the end of the shot when the motion gets fast, the quality of LDOF is better; see the hand motion in the last frame for instance.

Fig. 10 shows a particularly challenging sequence, since there is extremely fast motion of very small structures like the ball or the hands, and at the same time there are highly repetitive structures due to the fences in the background, which are problematic for the descriptor matching. Some persisting outliers can be seen in some frames. The latter problem gets worse at the end of the sequence after the camera has zoomed into the scene. Moreover, there is almost no structure on the tennis court, which can lead to arbitrary flow estimates in these areas.

LDOF performs quite well on this sequence, capturing the motion of the ball and the limbs most of the time. The racket is missed more often, which is also because it is partially transparent. The last three frames in this figure show the most typical limitations of the method that appear in this sequence: missed motion of the racket due to its changing appearance, false estimates near the fences due to too many false descriptor matches, and false motion of the tennis court due to missing structures and the moving shadow.

A sequence with a jumping monkey is shown in Fig. 11. Due to bad lighting conditions the image contrast is low and structures are blurred. This leads to several outliers in the window area, where mostly line structures are available, which are ambiguous. The actual motion of the jumping monkey and of the person defending the food is estimated quite well.

6 Discussion

In this paper we have presented a solution to the inherent problem of current state-of-the-art optical flow estimation methods to estimate large motions of small structures. This has been achieved by integrating correspondences from descriptor matching into a variational approach. While these

[2] The area along the right image boundary is a different story, since large parts disappear from the image.

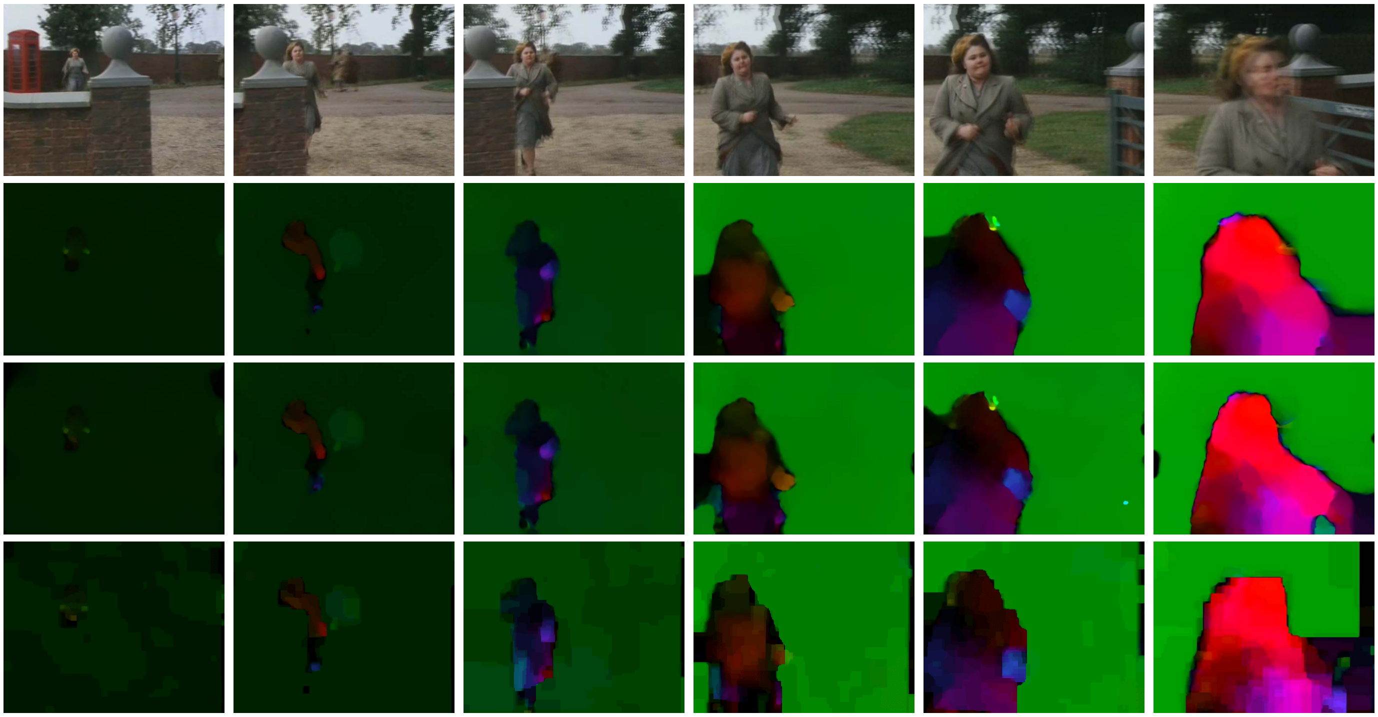

Fig. 8. Shot from *Miss Marple: A pocket full of rye*. **First row:** Input pairs overlaid to visualize the motion. **Second row:** Classic warping [11]. **Third row:** Large displacement optical flow (LDOF). **Fourth Row:** SIFT flow [23]. The hand in the last frame actually moves to the left, which is only captured by LDOF.

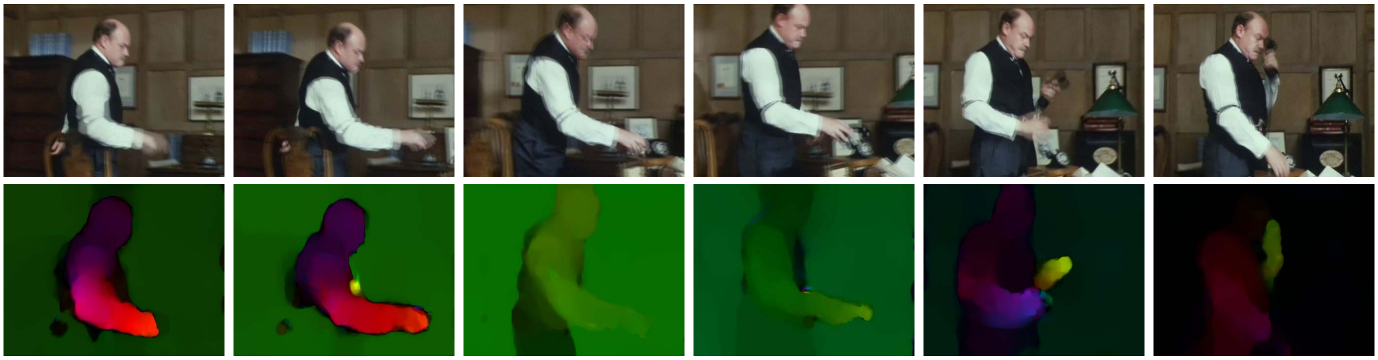

Fig. 9. Shot from *Miss Marple: A pocket full of rye*. **Top row:** Input pairs overlaid to visualize the motion. **Bottom row:** Flow fields estimated with LDOF.

correspondences are not intended to improve the accuracy of the approach, they support the coarse-to-fine warping strategy in avoiding local minima. As we have shown, both concepts towards better minima are complementary as they correspond to two different ways to split the non-convex optimization problem into a series of tractable subproblems.

We see applications of this technique particularly in action recognition and tracking. In action recognition, the new ability to estimate the optical flow of the most prominent, fast motions should help to better exploit the motion information in the video signal. For instance, the histogram of optical flow (HOF) features in [19] could directly benefit from our method. Tracking is traditionally based on descriptor matching in a local search window. We expect that the combination of local descriptor matching with a spatial regularity constraint will add robustness to tracking. Moreover, the subpixel accuracy should reduce the problem of drift inherent to tracking. Issues like temporal consistency and realtime performance still have to be dealt with, though. The fact that GPU implementations of classic warping techniques [36] already achieve 30fps on VGA resolution indicates that realtime performance is also feasible with the present approach. The additionally required descriptor matching fits very well to parallel hardware.

Apart from these traditional motion applications, the ability to estimate large displacements allows to reach out for more general matching problems. Works like [4] and [22] have demonstrated the potential of dense correspondence fields in recognition tasks. A variant of our technique might be applied in this field as well.

From a more technical point of view, the combination

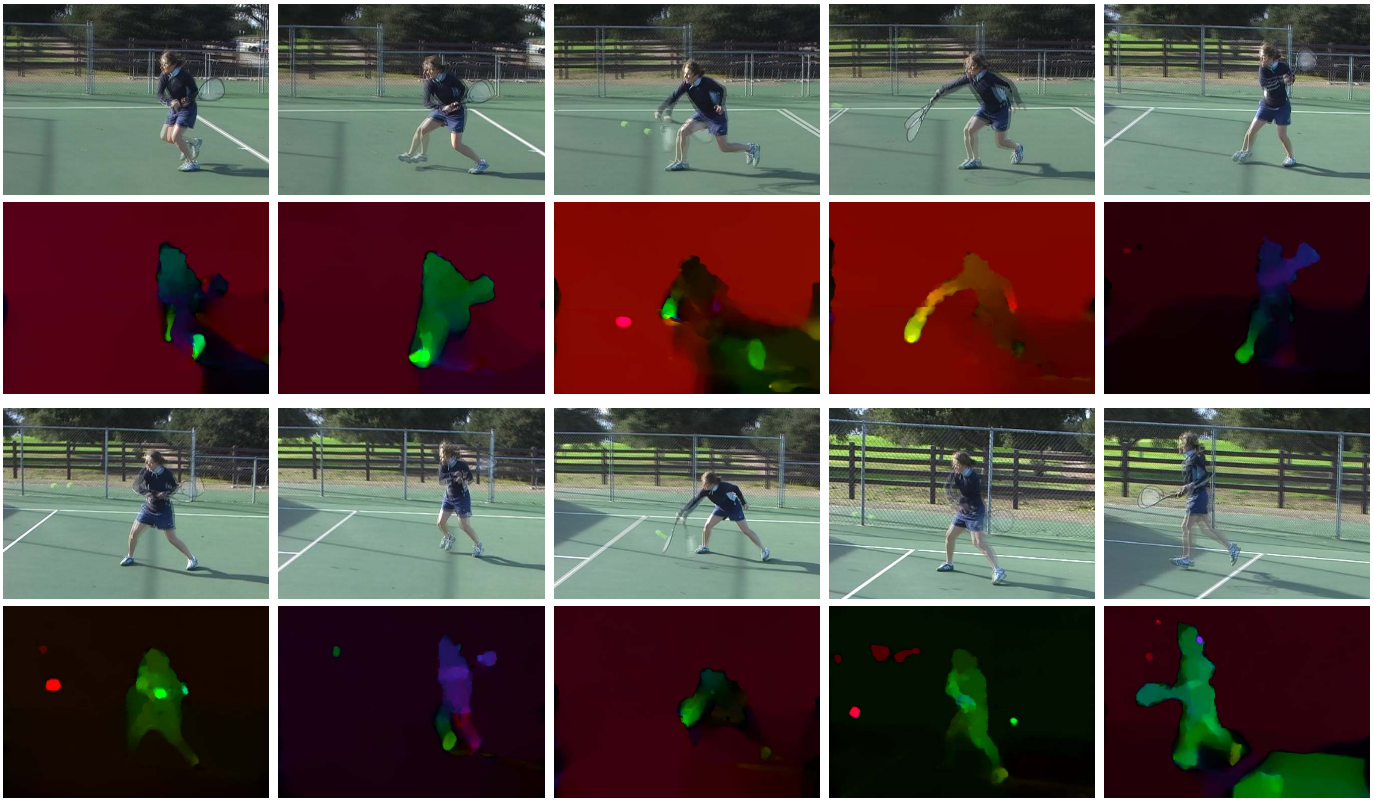

Fig. 10. Tennis sequence taken with a standard consumer camera. The motions of the legs, arms, the racket, and the ball are particularly challenging. Moreover, the fences in the background with their repetitive structures cause many errors in the descriptor matching. The flow fields have been estimated using HOG descriptors. The last three frames were picked to show some remaining limitations.

of discrete matching with a continuous, variational approach presented here can be investigated in more general settings. For instance, one could imagine to replace the simple nearest neighbor matching by a more efficient variant of the belief propagation approach from [29]. This one would not need to be particularly precise and could be computed on a coarser grid as the high precision job is done by the variational optimization.

Appendix

Rather than integrating only the best match, the K best matches can be considered by redefining the corresponding energy term as:

$$E_{\text{match}}(\mathbf{w}) = \int \sum_{i=1}^{K} \delta_i(\mathbf{x})\rho_i(\mathbf{x})\Psi\left(|\mathbf{w}(\mathbf{x}) - \mathbf{w}_i(\mathbf{x})|^2\right)\, d\mathbf{x}, \tag{15}$$

where each of the K hypotheses $\mathbf{w}_i(\mathbf{x})$ is weighted by its matching score $\rho_i(\mathbf{x})$. The advantage of integrating multiple hypotheses is the reduced number of missed correspondences. This plays a role when tracking extremely small structures like a tennis ball, which might be covered only by a single descriptor. A denser sampling of descriptors, however, can usually deal with this problem as well.

The disadvantage of integrating multiple hypotheses, besides the increased computational cost, is the significantly increased amount of mismatches that have to be sorted out by the variational approach. The robust function $\Psi(s^2) = \sqrt{s^2 + 0.001^2}$, which is related to the median of the data, can only deal with a limited outlier to inlier ratio. As a remedy, one could think of using non-convex error norms, but this leads to a harder optimization problem, which cannot be solved properly in the variational setting.

References

[1] L. Alvarez, J. Weickert, and J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision*, 39(1):41–56, Aug. 2000.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: an empirical evaluation. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2009.

[3] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. International Conference on Computer Vision*, 2007.

[4] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2005.

[5] A. Berg and J. Malik. Geometric blur for template matching. In *Proc. International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 607–614, 2001.

[6] M. J. Black and P. Anandan. Robust dynamic motion estimation over time. In *Proc. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 292–302, Maui, HI, June 1991. IEEE Computer Society Press.

[7] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, Jan. 1996.

[8] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, MA, 1987.

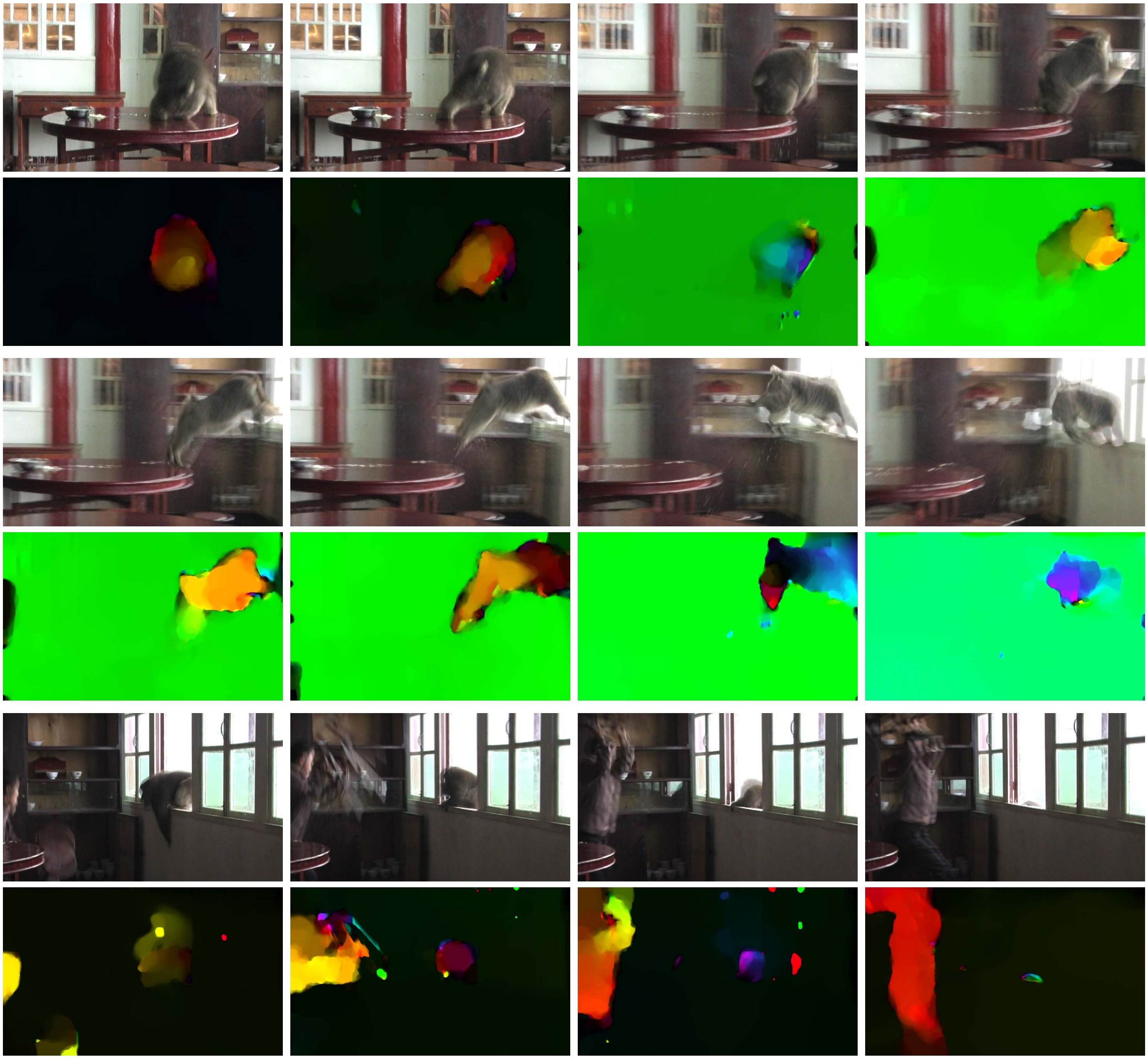

Fig. 11. Monkey sequence taken with a standard consumer camera. Due to bad lighting conditions, the image quality is rather low. The fast motion of the jumping monkey and the person defending the food is well captured, even though there are some occasional mismatches in the background.

[9] T. Brox. *From Pixels to Regions: Partial Differential Equations in Image Analysis*. PhD thesis, Faculty of Mathematics and Computer Science, Saarland University, Germany, Apr. 2005.

[10] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2009.

[11] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3024 of *LNCS*, pages 25–36. Springer, May 2004.

[12] A. Bruhn and J. Weickert. Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *Proc. 10th International Conference on Computer Vision*, pages 749–755. IEEE Computer Society Press, Beijing, China, Oct. 2005.

[13] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2–3):114–141, 2003.

[14] I. Cohen. Nonlinear variational method for optical flow computation. In *Proc. Eighth Scandinavian Conference on Image Analysis*, volume 1, pages 523–530, Tromsø, Norway, May 1993.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[16] T. Darrell, P. Indyk, and G. Shakhnarovich, editors. *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[17] P. Heas, E. Mémin, N. Papadakis, and A. Szantai. Layered estmation of atmospheric mesoscale dynamics from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):4087–4104, 2007.

[18] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2008.

[20] C. Lei and Y.-H. Yang. Optical flow estimation on coarse-to-fine region-trees using discrete optimization. In *Proc. International Conference on Computer Vision*, 2009.

[21] V. Lempitsky, S. Roth, and C. Rother. Discrete-continuous optimization for optical flow estimation. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2008.

[22] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2009.

[23] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: dense correspondence across different scenes. In *Proc. European Conference on Computer Vision*, volume 5304 of *LNCS*, pages 28–42. Springer, 2008.

[24] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[25] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. Seventh International Joint Conference on Artificial Intelligence*, pages 674–679, Vancouver, Canada, Aug. 1981.

[26] E. Mémin and P. Pérez. Dense estimation and object-based segmentation of the optical flow with robust techniques. *IEEE Transactions on Image Processing*, 7(5):703–719, May 1998.

[27] Y. Mileva, A. Bruhn, and J. Weickert. Illumination-invariant variational optical flow with photometric invariants. In *Pattern Recognition - Proc. DAGM*, volume 4713 of *LNCS*, pages 152–162. Springer, 2007.

[28] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67:141–158, 2006.

[29] A. Shekhovtsov, I. Kovtun, and V. V. Hlaváč. Efficient MRF deformation model for non-rigid image matching. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2007.

[30] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[31] F. Steinbruecker, T. Pock, and D. Cremers. Large displacement optical flow computation without warping. In *Proc. International Conference on Computer Vision*, 2009.

[32] D. Sun, S. Roth, J. P. Lewis, and M. J. Black. Learning optical flow. In *Proc. European Conference on Computer Vision*, volume 5304 of *LNCS*, pages 83–87. Springer, 2008.

[33] J. Weber and J. Malik. Robust computation of optical flow in a multi-scale differential framework. *International Journal of Computer Vision*, 14:67–81, 1995.

[34] A. Wedel, T. Pock, C. Zach, D. Cremers, and H. Bischof. An improved algorithm for TV-L1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, volume 5604 of *LNCS*, pages 23–45. Springer, 2009.

[35] J. Wills and S. Belongie. A feature based method for determining dense long range correspondences. In *Proc. European Conference on Computer Vision*, volume 3023 of *LNCS*, pages 170–182. Springer, 2004.

[36] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Pattern Recognition - Proc. DAGM*, volume 4713 of *LNCS*, pages 214–223. Springer, 2007.

[37] H. Zimmer, A. Bruhn, J. Weickert, L. Valgaerts, A. Salgado, B. Rosenhahn, and H.-P. Seidel. Complementary optic flow. In *Proc. EMMCVPR*, volume 5681 of *LNCS*, pages 207–220. Springer, 2009.

Thomas Brox received his Ph.D. degree in computer science from the Saarland University, Germany in 2005. Subsequently, he spent two years as a postdoctoral researcher at the University of Bonn, Germany, and one year as a temporary faculty member at the University of Dresden, Germany. He is currently a postdoctoral fellow at U.C. Berkeley. His research interest is in computer vision with special focus on video analysis, particularly optical flow estimation, motion segmentation, learning and detection in videos. In 2004, he received the Longuet-Higgins Best Paper Award at ECCV for his work on optical flow estimation.

Jitendra Malik received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1980 and the PhD degree in Computer Science from Stanford University in 1985. In January 1986, he joined the University of California at Berkeley, where he is currently the Arthur J. Chick Professor in the Department of Electrical Engineering and Computer Sciences. He is also on the faculty of the Cognitive Science and Vision Science groups. He serves on the advisory board of Microsoft Research India, and on the Governing Body of IIIT Bangalore. His research interests are in computer vision, computational modeling of human vision, and analysis of biological images. His work has spanned a range of topics in vision including image segmentation, perceptual grouping, texture, stereopsis, and object recognition with applications to image based modeling and rendering in computer graphics, intelligent vehicle highway systems, and biological image analysis. He has authored or co-authored more than 150 research papers on these topics, and graduated 26 PhD students who occupy prominent places in academia and industry. According to Google Scholar, five of his papers have received more than a thousand citations each. He received the gold medal for the best graduating student in Electrical Engineering from IIT Kanpur in 1980 and a Presidential Young Investigator Award in 1989. At U.C. Berkeley, he was selected for the Diane S. McEntyre Award for Excellence in Teaching in 2000 and a Miller Research Professorship in 2001. He received the Distinguished Alumnus Award from IIT Kanpur in 2008 and was awarded the Longuet-Higgins Prize for a contribution that has stood the test of time twice, in 2007 and in 2008. He is a fellow of the IEEE and the ACM.