Large Displacement Optical Flow*

Thomas Brox[1] Christoph Bregler[2] Jitendra Malik[1]

[1]University of California, Berkeley
Berkeley, CA, 94720, USA
{brox,malik}@eecs.berkeley.edu

[2]Courant Institute, New York University
New York, NY, 10003, USA
bregler@courant.nyu.edu

Abstract

The literature currently provides two ways to establish point correspondences between images with moving objects. On one side, there are energy minimization methods that yield very accurate, dense flow fields, but fail as displacements get too large. On the other side, there is descriptor matching that allows for large displacements, but correspondences are very sparse, have limited accuracy, and due to missing regularity constraints there are many outliers. In this paper we propose a method that can combine the advantages of both matching strategies. A region hierarchy is established for both images. Descriptor matching on these regions provides a sparse set of hypotheses for correspondences. These are integrated into a variational approach and guide the local optimization to large displacement solutions. The variational optimization selects among the hypotheses and provides dense and subpixel accurate estimates, making use of geometric constraints and all available image information.

1. Introduction

Optical flow estimation has been declared as a solved problem several times. For restricted cases this is true, but in more general cases, we are still far from a satisfactory solution. For instance estimating a dense flow field of people with fast limb motions cannot yet be achieved reliably with state-of-the-art techniques. This is of importance for many applications, like long range tracking, motion segmentation, or flow based action recognition techniques [5, 7].

Most contemporary optical flow techniques are based on two important ingredients, the energy minimization framework of Horn and Schunck [6], and the concept of coarse-to-fine image warping introduced by Lucas and Kanade [10] to overcome large displacements. Both approaches have been extended by robust statistics, which allow the treatment of outliers in either the matching or the smoothness assumption, particularly due to occlusions or motion discontinuities [3, 14]. The technique in [4] further introduced gradient constancy as a constraint which is robust to illumination changes and proposed a numerical scheme that allows for a very high accuracy, provided the displacements are not too large.

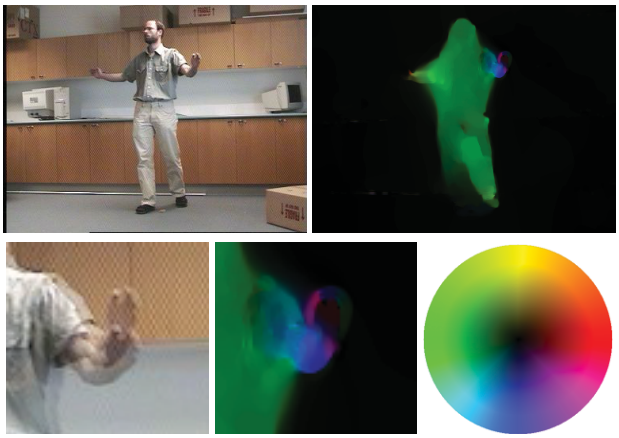

Figure 1. **Top row:** Image of a sequence where the person is stepping forward and moving his hands. The optical flow estimated with the method from [4] is quite accurate for the main body and the legs, but the hands are not accurately captured. **Bottom row, left:** Overlay of two successive frames showing the motion of one of the hands. **Center:** The arm motion is still good but the hand has a smaller scale than its displacement leading to a local minimum. **Right:** Color map used to visualize flow fields in this paper. Smaller vectors are darker and color indicates the direction.

The reason why differential techniques can deal with displacements larger than a few pixels at all is that they initialize the flow by estimates from coarser image scales, where displacements are small enough to be estimated by local optimization. Unfortunately, the downsampling not only smoothes the way to the global optimum, but also removes information that may be vital for establishing the correct matches. Consequently, the method cannot refine the flow of structures that are smaller than their displacement, simply because the structure is smoothed away just at the level when its flow is small enough to be estimated in the variational setting. The resulting flow is then close to the motion of the larger scale structure. This still works well if the motion varies smoothly with the scale of the structures, and even precise 3D reconstruction of buildings becomes possible [16]. Figure 1, however, shows an example, where the hand motion is not estimated correctly because the hand

*This work was funded by the German Academic Exchange Service (DAAD) and the ONR-MURI program.

is smaller than its displacement relative to the motion of the larger scale structure in the background. Such cases are very common with articulated objects.

If one is interested only in very few correspondences, descriptor matching is a widespread methodology to estimate arbitrarily large displacement vectors. Only a few points are selected for matching. Selected points should have good discriminative properties and there should be a high probability that the same point is selected in both images [17]. Quite some effort is put into the descriptors of the keypoints such that they are invariant to likely transformations of the surrounding patches. Due to their small number and their informative descriptors, *e.g*. SIFT [9], keypoints can be matched globally using a nearest neighbor criterion. In return, other disadvantages are present. Firstly, there is no geometric relationship per se enforced between matched keypoints. A counterpart to the smoothness assumption in optical flow is missing. Thus, outliers are very likely to appear. Secondly, correspondences are very sparse. Turning the sparse set into a dense flow field by interpolation leads to very inaccurate results missing most of the details.

In some applications, the dense 2D matching problem can be circumvented by making use of specific assumptions. If the scene is static and all motion in the image is due to the camera, the problem can be simplified by estimating the epipolar geometry from very few correspondences (established, *e.g*., by descriptor matching and some outlier removal procedure such as RANSAC) and then converting the 2D optical flow problem into a 1D disparity estimation problem. While the complexity of combinatorial optimization including geometric constraints in 2D is exponential, it becomes polynomial for some 1D problems. Consequently, large displacements are much less of a problem in typical stereo or structure-from-motion tasks, where dense disparity maps can be estimated via graph cut methods or similar techniques.

Unfortunately, this does not work any more as soon as objects besides the observer are moving. If the focus is on drawing information from the object motion rather than its static shape, there is no way around optical flow estimation, and although the image motion caused by moving objects in the scene is usually much smaller than that caused by a moving camera, displacements can still be too large for contemporary methods. This holds especially true as it is difficult to separate the egomotion of the camera from the object motion as long as both are not known.

For this reason we elaborate in the present paper on optical flow estimation with large displacements. The main idea is to direct a variational technique using correspondences from sparse descriptor matching. This aims at avoiding the local optimization to get stuck in a local minimum underestimating the true flow.

A recent work called SIFT Flow goes a step even further and tries to establish dense correspondences between different scenes [8]. The work is related to ours in the sense that rich descriptors are used in combination with geometric regularization. An approximative discrete optimization method from [15] is used to achieve this goal. The problem of this method in the context of motion estimation is due to the bad localization of the SIFT descriptor. Another strongly related work is the one by Wills and Belongie which allows for large displacements by using edge correspondences in a thin-plate-spline model [18].

In principle, any sparse matching technique can be used to find initial matches. However, it is important that the descriptor matching establishes correspondences also for the smaller scale structures missed by the coarse-to-fine optical flow. Here we propose to use regions from a hierarchical segmentation of the image. This has several advantages. Firstly, regions are more likely to coincide with separately moving structures than commonly used corners or blobs. Secondly, regions allow for estimates of affine patch deformations. Additionally, the hierarchical segmentation provides a good coverage of the whole image. This avoids missing some moving parts because there is no region detected in the area. There is another region-based detector [13], which has the first two properties but does not provide a hierarchy of regions. Another reasonable strategy is to enforce consistent segmentations between frames as suggested in [20].

An important issue is the combination of the keypoint matches and the raw image data within the variational approach. The straightforward way to initialize the variational method with the interpolated keypoint matches gives large influence to outliers. Moreover, it raises the question of which scale to initialize the variational method. The optimum scale is likely to vary from image to image. Therefore, we integrate the keypoint correspondences directly into the variational approach. This allows us to make use of all the image information (not only the keypoints) already at coarse levels, and smoothly scales down the influence of the keypoints as the grid gets finer. Moreover, we integrate multiple matching hypotheses into the variational energy. This allows us to postpone an important hard decision, namely which particular candidate region is the best match, to the variational optimization where geometric constraints are available. Thanks to this formulation, outliers are treated in a proper way, without the need to tune threshold parameters.

2. Region matching

2.1. Region computation

For creating regions in the image, we rely on the segmentation method proposed in Arbelaez et al. [1]. The segmentation is based on the boundary detector *gPb* from [11]. The

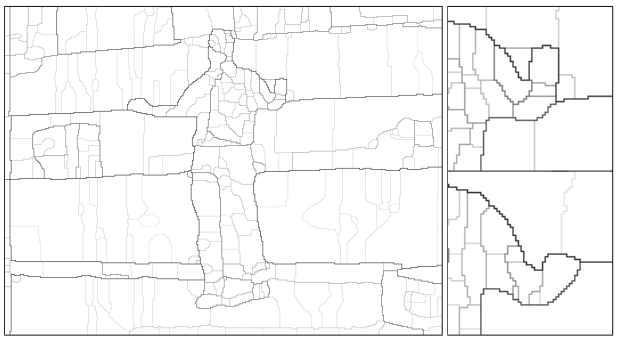

Figure 2. **Left:** Segmentation of an image. A region hierarchy is obtained by successively splitting regions at an edge of certain relevance. Dark edges are inserted first. **Right:** Zoom into the hand region of two successive images.

advantage of this boundary detector over simple edge detection is that it takes texture into account. Boundaries due to repetitive structures are damped whereas strong changes in texture create additional boundaries. Consequently, boundaries are more likely to correspond to objects or parts of objects. This is beneficial for our task, as it increases the stability of the regions to be matched.

The method returns a boundary map $g(\mathbf{x})$ as shown in Fig. 2. Strong edges correspond to more likely object boundaries. It further returns a hierarchy of regions created from this map. Regions with weak edges are merged first, while separations due to strong edges persist for many levels in the hierarchy. We generally take the regions from all the levels in the hierarchy into account. From the regions of the first image, however, we only keep the most stable ones, i.e., those which exist in at least 5 levels of the hierarchy. Unstable regions are usually arbitrary subparts of large regions. They are likely to change their shape between images. We also ignore extremely small regions (with less than 50 pixels) from both images. These regions are usually too small to build a descriptor discriminative enough for reliable matching.

2.2. Region descriptor and matching

To each region we fit an ellipse and normalize the area around the centroid of each region to a 32×32 patch. The normalized patch then serves as the basis for a descriptor.

We build two descriptors S and C in each region. S consists of 16 orientation histograms with 8 bins, like in SIFT [9]. C comprises the mean RGB color of the same 16 subparts as the SIFT descriptor. While the orientation histograms consider the whole patch to take also the shape of the region into account, the color descriptor is computed only from parts of the patch that belong to the region.

Correspondences between regions are computed by nearest neighbor matching. We compute the Euclidean distances of both descriptors separately and normalize them by the

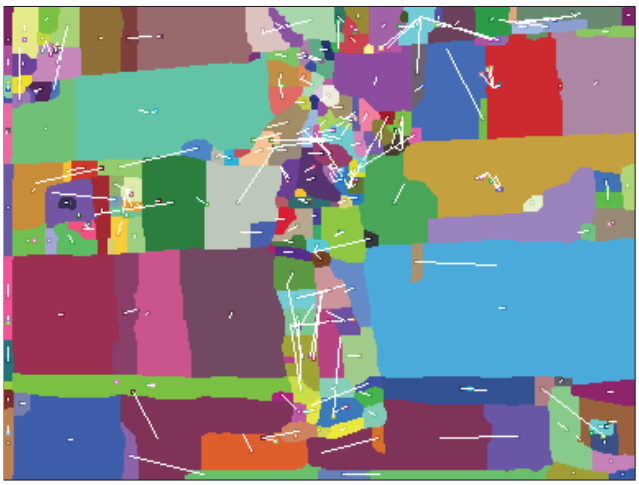

Figure 3. Displacement vectors of the matched regions drawn at their centroids. Many matches are good, but there are also outliers from regions that are not descriptive enough or their counterpart in the other image is missing.

sum over all distances:

$$\begin{aligned} d^2(S_i, S_j) &= \frac{\|S_i - S_j\|_2^2}{\frac{1}{N}|\sum_{k,l} \|S_k - S_l\|_2^2} \\ d^2(C_i, C_j) &= \frac{\|C_i - C_j\|_2^2}{\frac{1}{N}|\sum_{k,l} \|C_k - C_l\|_2^2}, \end{aligned} \tag{1}$$

where N is the total number of combinations i, j. This normalization allows to combine the distances such that both parts in average have equal influence:

$$d^2(i,j) = \frac{1}{2}(d^2(S_i, S_j) + d^2(C_i, C_j)) \tag{2}$$

We can exclude potential pairs by adding high costs to their distance. We do this for correspondences with a displacement larger than 15% of the image size or with a change in scale that is larger than factor 3. Depending on the needs of the application, these numbers can be adapted. Smaller values obviously produce fewer false matches, but restrict the allowed image transformations.

2.3. Hypotheses refinement by deformed patches

Fig. 3 demonstrates successful matching of many regions, but also reveals outliers. This is not surprising as some of the regions are quite small and not very descriptive. Moreover, the affine transformation estimated from the region shape is not always correct as the extracted regions may not be exactly the same in both images. Finally, the above descriptors are well suited to establish a ranking of potential matches, but the descriptor distance often performs badly when used as a confidence measure since good matches and bad matches have very similar distances.

Rather than deciding on a fixed correspondence at each keypoint, which could possibly be an outlier, we propose

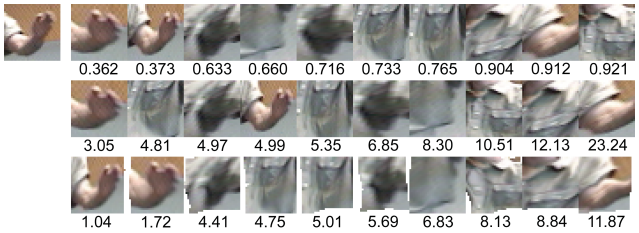

Figure 4. Nearest neighbors and their distances using different descriptors. **Top:** SIFT and color. **Center:** Patch within region. **Bottom:** Patch within region after distortion correction.

to integrate several potential correspondences into the variational approach. For this purpose, a good confidence measure is of great importance. We found that the distance of patches globally separates good and bad matches much better than the above descriptors. The main problem with direct patch comparison (classical block matching) is the sensitivity to small shifts or deformations. With the deformation corrected, the Euclidean distance of patches is very informative, particularly when considering only pixels from within the region[1].

The optimum shift and deformation needed to match two patches can be estimated by minimizing the following cost function:

$$E(u,v) = \int (P_2(x+u, y+v) - P_1(x,y))^2 \, dxdy + \alpha \int (|\nabla u|^2 + |\nabla v|^2) \, dxdy, \tag{3}$$

where P_1 and P_2 are the two patches, $u(x,y), v(x,y)$ denotes the deformation field to be estimated, and $\alpha = 10000$ is a tuning parameter that steers the relative importance of the deformation smoothness. The energy is a non-linearized, large displacement version of the Horn and Schunck energy and sufficient for this purpose. The regularizer gets a very high weight in this case, as without regularization every patch can be made sufficiently similar to any other.

As the patches are very small and a simple quadratic regularizer is applied, the estimation is quite efficient. Nevertheless, it would be a computational burden to estimate the deformation for each region pair. To this end, we preselect the 10 nearest neighbors for each patch using the distance from the previous section and compute the deformation only for these candidates. The five nearest neighbors according to the patch distance are then integrated into the variational approach described in the next section. Each potential match $j = 1, ..., 5$ of a region i comes with a confidence

$$c_j(i) := \begin{cases} \frac{\bar{d}^2(i) - d^2(i,j)}{d^2(i,j)} & \bar{d}^2(i) > 0 \\ 0 & \text{else} \end{cases} \tag{4}$$

[1] In contrast to tracking and motion estimation, this probably does not hold for object class detection.

where $d^2(i,j)$ is the Euclidean distance between the two patches after deformation correction and $\bar{d}^2(i)$ is the average Euclidean distance among the 10 nearest neighbors. This measure takes the absolute fit as well as the descriptiveness into account. We restrict the distance to be computed only on patch positions within the region. Hence the changing background of a moving object part would not destroy similarity of a correct match.

Fig. 4 depicts the nearest neighbors of a sample region. Simple block matching is clearly inferior compared to SIFT and color because the high frequency information is not correctly aligned. However, computing distances on distortion corrected patches is advantageous for our task. Not only the ranking improves in this particular case, the distance is in general also more valuable as a confidence measure since it marks bad matches more clearly.

3. Variational flow

Although most of the correspondences in Fig. 3 are correct, the flow field derived from these by interpolation, as shown in Fig. 5, is far from being accurate. This is because we have a hard decision when selecting the nearest neighbor. Moreover, a lot of image information is neglected and substituted by a smoothness prior. In order to obtain a more accurate, dense flow field, we integrate the matching hypotheses into a variational approach, which combines them with local information from the raw image data and a smoothness prior.

3.1. Energy

The energy we optimize is similar to the one in [4] except for an additional data constraint that integrates the correspondence information:

$$\begin{aligned} E(\mathbf{w}(\mathbf{x})) &= \int \Psi\left(|I_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) - I_1(\mathbf{x})|^2\right) d\mathbf{x} \\ &+ \gamma \int \Psi\left(|\nabla I_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) - \nabla I_1(\mathbf{x})|^2\right) d\mathbf{x} \\ &+ \beta \sum_{j=1}^{5} \int \rho_j(\mathbf{x}) \, \Psi\Big((u(\mathbf{x}) - u_j(\mathbf{x}))^2 + (v(\mathbf{x}) - v_j(\mathbf{x}))^2\Big) d\mathbf{x} \\ &+ \alpha \int \Psi\left(|\nabla u(\mathbf{x})|^2 + |\nabla v(\mathbf{x})|^2 + g(\mathbf{x})^2\right) d\mathbf{x} \end{aligned} \tag{5}$$

Here, I_1 and I_2 are the two input images, $\mathbf{w} := (u, v)$ is the sought optical flow field, and $\mathbf{x} := (x, y)$ denotes a point in the image. $(u_j, v_j)(\mathbf{x})$ is one of the motion vectors derived at position $\mathbf{x}$ by region matching (j indexing the 5 nearest neighbors). If there is no correspondence at this position, $\rho_j(\mathbf{x}) = 0$. Otherwise, $\rho_j(\mathbf{x}) = c_j$, where c_j is the distance based confidence in (4). $\alpha = 100$, $\beta = 25$, and $\gamma = 5$ are tuning parameters, which steer the importance of smoothness, region correspondences, and gradient constancy, respectively.

Like in [4], we use the robust function $\Psi(s^2) = \sqrt{s^2 + 10^{-6}}$ in order to deal with outliers in the data as well as in the smoothness assumption. We also integrate the boundary map $g(\mathbf{x})$ from [1] (see Fig. 2) in order to avoid smoothing across strong region boundaries.

The robust function further reduces the influence of bad correspondences and leads to the selection of the most consistent match among the five nearest neighbors. Note that each potential match has its own robust function. Spatial consistency is enforced by the smoothness prior, which integrates correspondences from the neighborhood. Many good matches in the neighborhood will outnumber mismatches, which is not the case when using a squared error measure. With $\alpha = 0$ the optimum result would simply be the weighted median of the hypotheses, but with $\alpha > 0$ additional matches from the surroundings are taken into account.

Rather than a straightforward three step procedure with (i) interpolation of the region correspondences, (ii) removal of outliers not fitting the interpolated flow field (iii) optical flow estimation initialized by the interpolated inlier correspondences, the above energy combines all three steps in a single optimization problem.

3.2. Minimization

The energy is non-convex and can only be optimized locally. We can compute the Euler-Lagrange equations, which state a necessary condition for a local optimum:

$$\begin{aligned}
&\Psi'\left(I_z^2\right) I_z I_x + \gamma \Psi'\left(I_{xz}^2 + I_{yz}^2\right)\left(I_{xx} I_{xz} + I_{xy} I_{yz}\right) \\
&\quad + \beta \textstyle\sum_j \rho_j \Psi'\left((u-u_j)^2 + (v-v_j)^2\right)(u-u_j) \\
&\quad - \alpha \operatorname{div}\left(\Psi'\left(|\nabla u|^2 + |\nabla v|^2 + g(\mathbf{x})^2\right)\nabla u\right) = 0 \\
&\Psi'\left(I_z^2\right) I_z I_y + \gamma \Psi'\left(I_{xz}^2 + I_{yz}^2\right)\left(I_{xy} I_{xz} + I_{yy} I_{yz}\right) \\
&\quad + \beta \textstyle\sum_j \rho_j \Psi'\left((u-u_j)^2 + (v-v_j)^2\right)(v-v_j) \\
&\quad - \alpha \operatorname{div}\left(\Psi'\left(|\nabla u|^2 + |\nabla v|^2 + g(\mathbf{x})^2\right)\nabla v\right) = 0,
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
I_x &:= \partial_x I_2(\mathbf{x}+\mathbf{w}) & I_{xy} &:= \partial_{xy} I_2(\mathbf{x}+\mathbf{w}) \\
I_y &:= \partial_y I_2(\mathbf{x}+\mathbf{w}) & I_{yy} &:= \partial_{yy} I_2(\mathbf{x}+\mathbf{w}) \\
I_z &:= I_2(\mathbf{x}+\mathbf{w}) - I_1(\mathbf{x}) & I_{xz} &:= \partial_x I_z \\
I_{xx} &:= \partial_{xx} I_2(\mathbf{x}+\mathbf{w}) & I_{yz} &:= \partial_y I_z.
\end{aligned} \tag{7}$$

Although we have the region correspondences involved in these equations, their influence would be too local to effectively drive a large displacement solution. However, we can make use of the same coarse-to-fine strategy as used in optical flow warping schemes. This has two effects. Firstly, downsampled large scale structures drive the optical flow to a large displacement solution. Secondly, the influence of region correspondences is much larger at coarser levels as they cover larger parts of the discrete domain. $\rho(\mathbf{x}) \neq 0$ for the same number of grid points, but the total number of grid points at coarser levels is much smaller. As a consequence, they dominate the optical flow at coarse levels, pushing the local optimization into the right direction. At finer levels, their influence decreases (and is actually zero in the true continuous case). While correct matches will be in line with the optical flow, outliers will be outnumbered by the growing number of grid points indicating a different flow field.

We can use the same nested fixed point iterations as proposed in [4] to solve (6). We initialize $\mathbf{w}^0 := (0,0)$ at the coarsest grid and iteratively compute updates $\mathbf{w}^{k+1} = \mathbf{w}^k + \mathbf{dw}^k$, where $\mathbf{dw}^k := (du^k, dv^k)$ is the solution of

$$\begin{aligned}
0 &= \Psi'_1 I_x^k (I_z^k + I_x^k du^k + I_y^k dv^k) \\
&\quad + \beta \textstyle\sum_j \rho_j \Psi'_{2,j}(u - u_j) - \alpha \operatorname{div}\left(\Psi'_3 \nabla(u^k + du^k)\right) \\
0 &= \Psi'_1 I_y^k (I_z^k + I_x^k du^k + I_y^k dv^k) \\
&\quad + \beta \textstyle\sum_j \rho_j \Psi'_{2,j}(v - v_j) - \alpha \operatorname{div}\left(\Psi'_3 \nabla(v^k + dv^k)\right)
\end{aligned} \tag{8}$$

with

$$\begin{aligned}
\Psi'_1 &:= \Psi'\left((I_z^k + I_x^k du^k + I_y^k dv^k)^2\right) \\
\Psi'_{2,j} &:= \Psi'\left((u^k + du^k - u_j)^2 + (v^k + dv^k - v_j)^2\right) \\
\Psi'_3 &:= \Psi'\left(|\nabla(u^k + du^k)|^2 + |\nabla(v^k + dv^k)|^2 + g^2\right).
\end{aligned} \tag{9}$$

We skipped the gradient constancy term in the notation to have shorter equations. The reader is referred to [4] for the gradient constancy part. In order to solve (8), an inner fixed point iteration over l is employed, where the robust functions in (9) 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.

4. Experiments

We evaluated the new method on several real images showing large displacements, particularly articulated motion of humans. Fig. 5 depicts results for the example from the previous sections. The fast motion of the person's right hand missed by current state-of-the-art optical flow is correctly captured when integrating point correspondences from descriptor matching. This clearly shows the improvement we aimed at. In areas without large displacements, we cannot expect the flow to be more accurate, since descriptor matching is not as precise as variational flow. However, the result is also not much spoiled by unprecise and bad matches. We quantitatively confirmed this by running [4] and the large displacement flow on five sequences of the Middlebury dataset with public ground truth [2]. There are no large displacements in any of these sequences. We optimized the parameters of both approaches but kept the

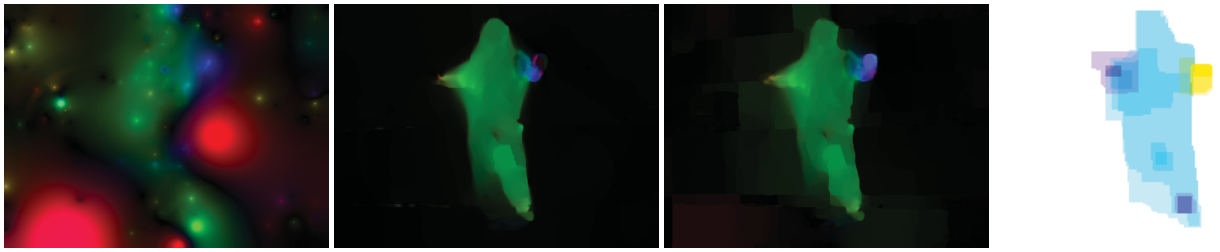

Figure 5. **Left:** Flow field obtained by interpolating the region correspondences (nearest neighbor). Accuracy is low and several outliers can be observed. **Center left:** Result with the optical flow method from [4]. The motion is mostly very accurate, but the hand motion is not captured well. **Center right:** Result with the proposed method. Most of the accuracy of the optical flow framework is preserved and the fast moving hands are captured as well. We see some degradations in the background due to outliers and too little structure to correct this. **Right:** Result of SIFT Flow [8] running the code provided by the authors. Since the histograms in SIFT lack good localization properties, the accuracy of the flow field is much lower.

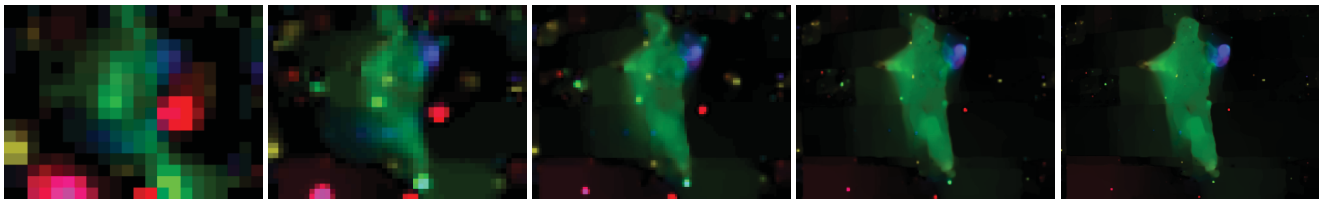

Figure 6. Evolving flow field from coarse (left) to fine (right). The region correspondences dominate the estimate at the beginning. Outliers are removed over time as more and more data from the image is taken into account.

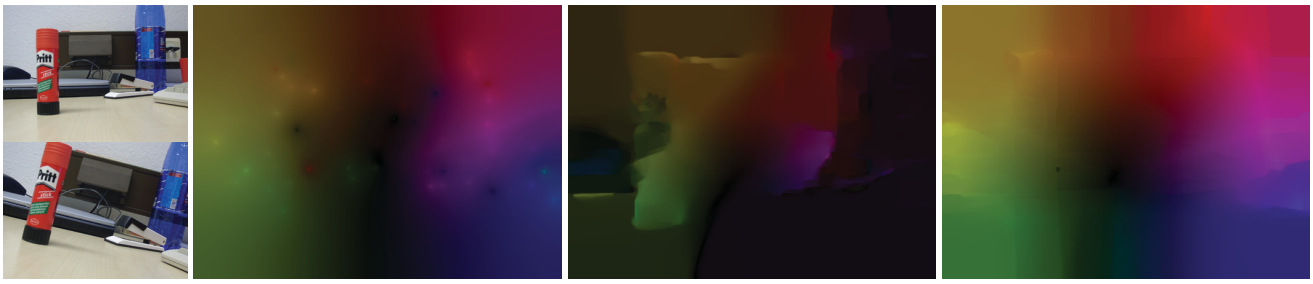

Figure 7. **Left:** Input images. The camera was rotated and moved into the scene. **Center left:** Interpolated region correspondences. **Center right:** Result with the optical flow method from [4]. Clearly, only the smaller displacements in the center and those of regions with appropriate scale can be estimated. **Right:** Result with the proposed method. Aside from the unstructured and occluded areas near the image boundaries, the flow field is estimated well.

parameter β (which steers the influence of the point correspondences) at the same value as in the other experiments. The average angular error of the large displacement version increased in average by 27%. This means it still yields a good accuracy while being able to capture larger motion.

Fig. 5 also demonstrates the huge improvement over descriptor matching succeeded by interpolation to derive a dense flow field. Clearly, the weakly descriptive information in the image aside of the keypoints should not be ignored when more than a few correspondences are needed. A comparison to [8] using their code indicates that we get a better localization of the motion, which is quite natural as [8] was designed to match between different scenes.

Fig. 6 shows the evolving flow field over multiple scales. In can be seen that the influence of wrong region matches decreases as the flow field includes more and more information from the image and geometrically inconsistent matches are ignored as outliers.

Fig. 7 depicts an experiment with a static scene and a moving camera. This problem would actually be better solved by estimating the fundamental matrix from few correspondences and then computing the disparity map with global combinatorial optimization. We show this experiment to demonstrate that good results can be obtained even without exploiting the knowledge of a static scene (which may not always be true in many realistic tasks). The flow in non-occluded areas is well estimated despite huge displacements. Neither classical optical flow nor interpolated descriptor matching can produce these results.

Another potential field of application of our technique is human motion analysis. Fig. 8 shows two frames from the HumanEva-II benchmark at Brown University. The original sequence was captured with a 120fps highspeed camera. We skipped four frames to simulate the 30fps of a consumer camera. Again we can see that the large motion of some body parts is missed with previous optical flow techniques,

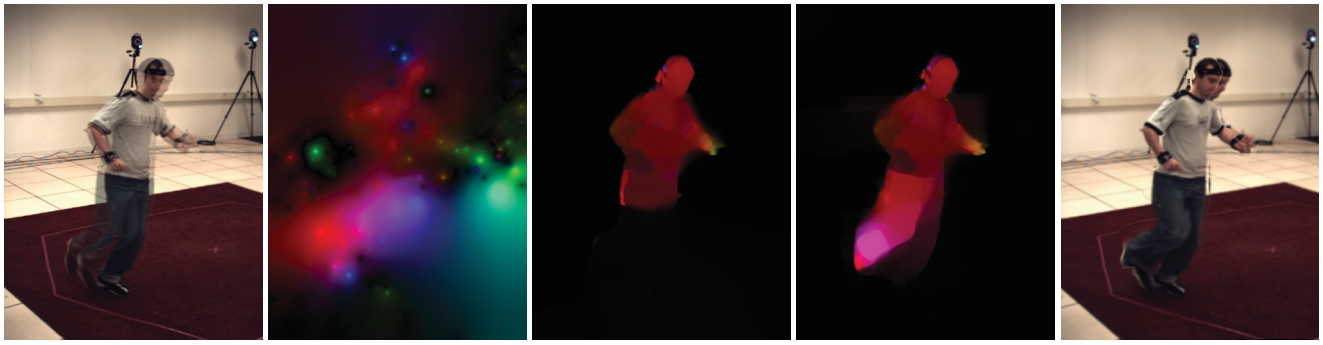

Figure 8. **Left:** Two overlaid images of a running person. The images are from the HumanEva-II benchmark on human tracking. **Center left:** Interpolated region correspondences. **Center:** Result with optical flow from [4]. The motion of the right leg is too fast to be captured by the coarse-to-fine scheme alone. **Center right:** Result with the proposed model. Region correspondences guide the optical flow towards the fast motion of the leg. **Right:** Image warped according to the estimated flow. The ideal result should look like the first image apart from occluded areas. The motion of the foot tip is underestimated, but the motion of the lower leg and the rest of the body is fine.

while it is captured much better when integrating descriptor matching. The warped image reveals that the motion of the foot tip is still underestimated, but the rest of the body including the lower leg and the arms is tracked correctly. The doubles near object boundaries are due to occlusion and indicate the correct filling of the background's zero motion.

Finally, Fig. 9-10 show results from a tennis sequence. The entire sequence and the corresponding flow is available in the supplementary material. The sequence was recorded with a 25fps hand held consumer camera and is very difficult due to very fast motion of the tennis player, little structure on the ground, and highly repetitive structures at the fence. The latter produce many outliers when matching regions. The video shows that most of the outliers are ignored in the course of variational optimization and also large parts of the fast motion is captured correctly. Jittering of the camera is indicated by the changing color in the background (showing changing motion directions). Even the motion of the ball is estimated in some frames. The motion of the racket and the hands is missed from time to time due to motion blur and weakly discriminative regions. Nevertheless, the results are very promising to serve as a cue in action recognition.

Computation of the flow for given segmentations took 37s on an Intel Xeon 2.33GHz for images of size 530×380 pixels. Most of the time is spent for the deformation of the patches and the variational flow, which is both potentially available in real-time using the GPU [19]. A GPU implementation of the segmentation takes 5s per frame.

5. Conclusions

We have shown that optical flow can benefit from sparse point correspondences from descriptor matching. The local optimization involved in optical flow methods fails to capture large motions even with coarse-to-fine strategies if small subparts move considerably faster than their surroundings. Point correspondences obtained from global nearest neighbor matching using strong descriptors can guide the local optimization to the correct large displacement. Conversely, we have also shown that weakly descriptive information, as is thrown away when selecting keypoints, contains valuable information and should not be ignored. The flow field obtained by exploiting all image information is much more accurate than the interpolated point correspondences. Moreover, outliers can be avoided by integrating multiple hypotheses into the variational approach and making use of the smoothness prior to select the most consistent one.

This work extends the applicability of optical flow to fields with larger displacements, particularly to tasks where large displacements are due to object rather than camera motion. We expect good results in action recognition when using the dense flow as a dynamic orientation feature correspondingly to orientation histograms in static image recognition. However, with larger displacements there also appear new challenges such as occlusions, which we mainly ignored here. Future works should transfer the rich knowledge on occlusion handling in disparity estimation to the more general field of optical flow.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: an empirical evaluation. *Proc. CVPR*, 2009.

[2] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Proc. ICCV*, 2007.

[3] 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, 1996.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. *Proc. ECCV*, Springer LNCS 3024, 25–36, 2004.

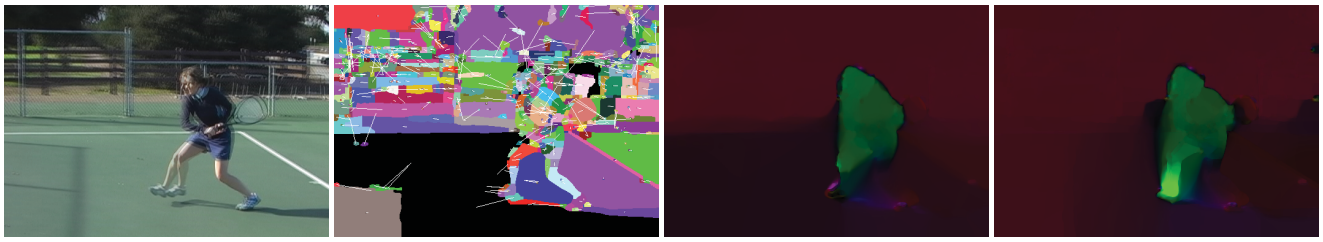

Figure 9. **Left:** Two overlaid images of a tennis player in action. **Center left:** Region correspondences. **Center right:** Result with optical flow from [4]. The motion of the right leg is too fast to be estimated. **Right:** The proposed method captures the motion of the leg.

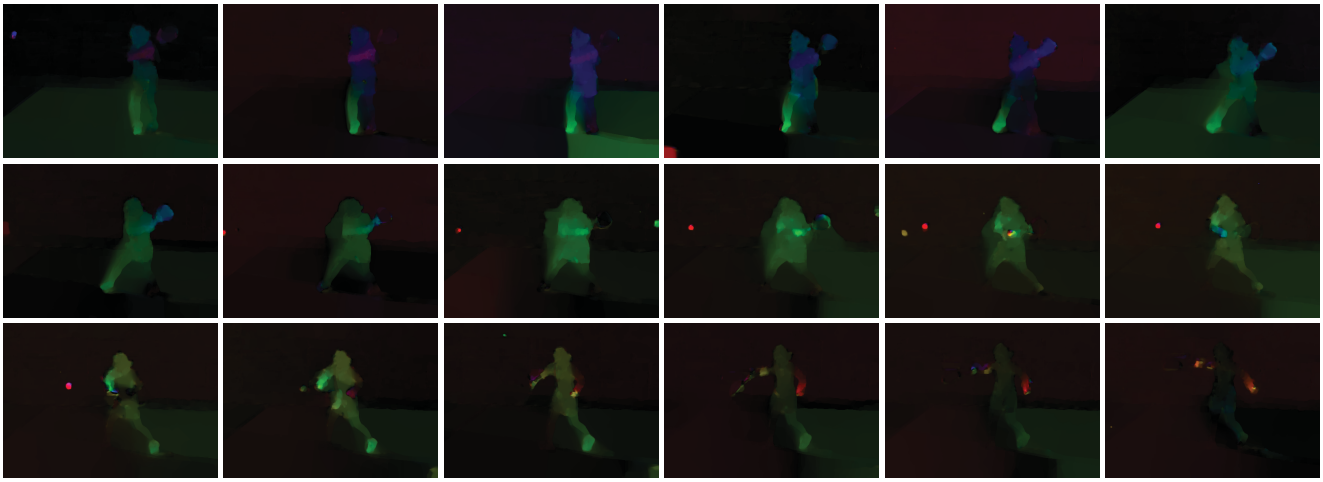

Figure 10. One figure of a tennis sequence obtained with a 25fps hand held consumer camera. Despite extremely fast movements most part of the interesting motion is correctly captured. Even the ball motion (red) is estimated here. The entire video is available as supplementary material.

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *Proc. ICCV*, 726–733, 2003.

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

[7] I. Laptev. *Local Spatio-Temporal Image Features for Motion Interpretation*. PhD thesis, Computational Vision and Active Perception Laboratory, KTH Stockholm, Sweden, 2004.

[8] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: dense correspondence across different scenes. *Proc. ECCV*, Springer LNCS 5304, 28–42, 2008.

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

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

[11] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. *Proc. CVPR*, 2008.

[12] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Proc. British Machine Vision Conference*, 2002.

[14] 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, 1998.

[15] A. Shekhovtsov, I. Kovtun, and V. V. Hlaváč. Efficient MRF deformation model for non-rigid image matching. *Proc. CVPR*, 2007.

[16] C. Strecha, R. Fransens, and L. Van Gool. A propabilistic approach to large displacement optical flow and occlusion detection. *Statistical Methods in Video Processing*, Springer LNCS 3247, 71–82, 2004.

[17] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[18] J. Wills and S. Belongie. A feature based method for determining dense long range correspondences. *Proc. ECCV*, Springer LNCS 3023, 170–182, 2004.

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

[20] C. L. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. *Proc. ICCV*, 2005.