SUPERPIXEL-BASED LARGE DISPLACEMENT OPTICAL FLOW

Haw-Shiuan Chang and Yu-Chiang Frank Wang

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

It has been a challenging task to estimate optical flow for videos in which either foreground or background exhibits remarkable motion information (i.e., large displacement), or those with insufficient resolution due to artifacts like motion blur or noise. We present a novel optical flow algorithm, which approaches the above problem as solving the task of energy minimization, which exploits image data and smoothness terms at the superpixel level. Our proposed method can be considered as an extended mean-shift algorithm, which advances color and gradient information of superpixels across consecutive frames with smoothness guarantees. Since we do not require assumptions of linearlization during optimization (as standard optical flow approaches do), we are able to alleviate local minimum problems and thus produce improved estimation results. Empirical results on the MPI-Sintel video dataset verify the effectiveness of our proposed method.

Index Terms— large displacement optical flow, superpixel, mean shift

1. INTRODUCTION

Optical flow has been widely applied in computer vision applications such as motion estimation, object segmentation, and video stabilization. Calculating optical flow across video frames can be considered as the task of extracting motion patterns between consecutive frames. In practice, videos with insufficient resolution or those being corrupted due to occlusion, motion blur, etc. noise would cause optical flow estimation error. Moreover, if foreground or background regions exhibit significant motion variations (i.e., *large displacement*), how to properly calculate the optical flow across video frames will be a very challenging task.

Originally proposed by Horn and Schunck [1], optical flow is calculated by a variational model which solves an optimization problem of data and smoothness terms. The data term typically requires the assumption of linearization for matching *local* image brightness or gradients between consecutive frames, and the smoothness term aims at preserving the spatial consistency of the resulting optical flow. However, if a video exhibits significant motion information, the above assumption would not be valid and thus the estimation turns into a nonlinear/non-convex optimization problem.

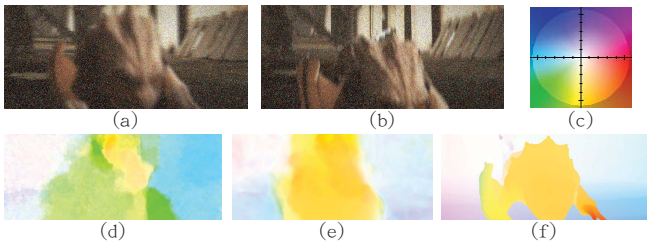

Fig. 1. Example of large displacement videos with insufficient resolution. (a)-(b) Consecutive frames, (c) color code for visualizing optical flow, (d) estimation of a standard coarse-to-fine model [3], (e) our result, and (f) ground truth optical flow.

To address the problem of large displacements during optical flow estimation, coarse-to-fine warping has been applied in [2, 3], which downsample the resolution of video frames for handling optical flow with significant motion. However, motion of articulated foreground or background regions with *smaller* scales would be disregarded after downgrading the resolution. Others proposed to utilize descriptor matching (e.g., SIFT) for addressing this problem [4]. To further improve the accuracy of the above matching scheme, Large Displacement Optical Flow (LDOF) [5] integrated both coarse-to-fine warping variational model and descriptor matching and promising result were reported in [5]. Nevertheless, descriptor matching might *not* be preferable for videos without sufficient resolution or with noise presented.

Instead of relying on variational models, superpixel-based approaches have been recently proposed in [6, 7] for improved estimation. This type of methods assumes the consistency of the optical flow estimated within each image segment (i.e., superpixel) and thus alleviates the issues of insufficient video resolution or noise presented. However, methods like [6, 7] did not consider videos with large displacement, and thus they cannot produce large displacement optical flow.

In this work, we particularly address optical flow estimation for videos with large motion (foreground or background) but with insufficient resolution due to motion blur or noise. As shown in Figure 1, traditional warping-based methods might not generalize well on such videos. Different from a recent work of [8] which addressed such problems for videos with large displacements by decoupling data and smoothness terms and performing pixel-level matching, we advocate the use of image superpixels for jointly optimizing both data and smoothness terms during optical flow estima-

tion. Since we do *not* require the assumption of linearization for our proposed model, the search space of our method is expanded and thus local minimum problem can be alleviated. In our experiments, we will verify that our method quantitatively and qualitatively outperforms state-of-the-art optical flow estimation approaches on such videos.

2. OUR PROPOSED METHOD

2.1. Problem Formulation

Traditional optical flow techniques are typically performed at the pixel level, and they cannot be easily extended to videos with large motion but with insufficient resolution. To extend the search range for the above cases while preserving image details, we propose to calculate superpixel-based optical flow from I_1 to I_2 via $\sum_{p\in i} I_2(\mathbf{x}_p) - I_1(\mathbf{x}_p - \mathbf{u}_i)$, where $\mathbf{u}_i$ is the estimated optical flow for the ith superpixel, and $\mathbf{x}_p$ is the location of pixel p. In our work, we utilize color, gradient, and spatial information of superpixels extracted at I_1 for estimating the optical flow. To be more precise, we define the energy function to be minimized as follows:

$$E = \sum_i \left(E_i^C(\mathbf{u}_i) + \sum_{j\in N_i^G} w_{(i,j)}^G E_{(i,j)}^G(\mathbf{u}_i) + \sum_{j\in N_i^S} w_{(i,j)}^S E_{(i,j)}^S(\mathbf{u}_i, \mathbf{u}_j) \right), \quad (1)$$

where N_i^G and N_i^S indicate the neighborhood sets of ith superpixel (discussed in Sections 2.1.2 and 2.1.3), and j denotes the index of neighboring superpixels of i. The energy terms E^C, E^G, and E^S preserve color, gradient and smoothness consistency, respectively. Parameters w^G and w^S balance the gradient and smoothness terms. In our work, we apply Turbopixels proposed by [9] for performing over-segmentation at each frame. As verified by [10], this allows one to obtain edge-preserving superpixels with similar sizes.

2.1.1. Color Energy Term

We first discuss the data term E_i^C in (1) for color consistency, which is defined as:

$$E_i^C(\mathbf{u}_i) = \sum_{p\in B_i^C} g(\|\mathbf{x}_p - (\mathbf{x}_i + \mathbf{u}_i)\|_2)\rho\left(\left\|\mathbf{I}_2(\mathbf{x}_p) - \mu_i^C\right\|_2\right), \quad (2)$$

where $\mathbf{x}_i$ is the 2D location of the center of the ith superpixel, and $\mathbf{u}_i$ is the estimated optical flow. We have μ_i^C as a three-dimensional vector, where each entry indicates the median value of the color channel R, G, or B for superpixel i at I_1. We have B_i^C as the set of pixels considered for the ith superpixel, and $\mathbf{I}_2(\mathbf{x}_p)$ representing the color information of pixel p within B_i^C at I_2.

In (2), a Gaussian function g relates the spatial information between each $\mathbf{x}_p$ within B_i^C and the estimated superpixel center $\mathbf{x}_i + \mathbf{u}_i$ at I_2. If a video is with insufficient resolution, superpixels with larger sizes will be produced (and thus a larger B_i^C), and this allows us to increase the search range during optical flow estimation.

The second term in (2) advances $\rho(x) = -exp(-x/\beta)$ as the penalty function. This function measures the similarity between μ_i^C of I_1 and the color information for pixel p in I_2. Its parameter β is calculated as the average difference between all μ_i^C and those of neighboring superpixels. It can be seen that, when calculating the color energy term, the Gaussian function g would suppress the color difference for pixels farther away from the superpixel center. As a result, estimation error due to color inconsistency would be alleviated.

2.1.2. Gradient Energy Term

As noted earlier, we over-segment each video frame by Turbopixels, which produces superpixels with similar sizes while preserving edge information. To preserve gradient consistency between each superpixel i and its neighbors (with index j) for calculating the optical flow $\mathbf{u}_i$, we determine the corresponding data energy term E^G in (1) as:

$$E_{(i,j)}^G(\mathbf{u}_i) = \sum_{p\in B_{(i,j)}^G} g(\|\mathbf{x}_p - (\mathbf{x}_{(i,j)} + \mathbf{u}_i)\|_2)\rho\left(\left\|\nabla\mathbf{I}_2(\mathbf{x}_p) - \mu_{(i,j)}^G\right\|_2\right). \quad (3)$$

In (3), $\mathbf{x}_p$ is the location of pixel p, and $\mathbf{x}_{(i,j)}$ denotes the center of the boundary between this superpixel and its neighbors at I_1. Similar to the color energy term defined in (2), $\mu_{(i,j)}^G$ indicates the median gradient information along this boundary, and $\nabla\mathbf{I}_2(\mathbf{x}_p)$ calculates the gradient of I_2 at $\mathbf{x}_p$. Finally, $B_{(i,j)}^G$ represents the pixel set whose size (in each dimension) is determined by the length of the associated boundary.

We note that E^G has the same form as E^C in (2), and the matching between gradient information via ρ is also weighted by a spatial Gaussian function g. For each video frame, we take the average gradient magnitude μ^G as the parameter β for the ρ function.

2.1.3. Smoothness Term

Finally, we define the smoothness term $E_{(i,j)}^S(\mathbf{u}_i, \mathbf{u}_j)$ when calculating the optical flow $\mathbf{u}_i$ and $\mathbf{u}_j$ for the ith superpixel and its neighbor j as follows:

$$\sum_{j\in N_i^S} w_{(i,j)}^S E_{(i,j)}^S(\mathbf{u}_i, \mathbf{u}_j) = \sum_{j\in N_i^S} w_{(i,j)}^S \|\mathbf{u}_i - \mathbf{u}_j\|_2^2. \quad (4)$$

Unlike the gradient term E^G which considers the neighboring superpixels connecting to the ith superpixel as N_i^G, N_i^S in (4) now represents a larger neighbor superpixel set and

contains the neighboring superpixels of N_i^G. This is to preserve the spatial smoothness consistency when calculating the superpixel-based optical flow (instead of enforcing the above consistency in locally neighboring ones).

We note that, when estimating the optical flow using the method of [11], nonlocal and anisotropic regularization on the smoothness term has been shown to improve the performance. By advancing superpixel-based optical flow with quadratically nonlocal and anisotropic regularization, better estimation results can be expected for noisy or blurred videos with large displacements, as verified later in Section 3.

2.2. Optimization

We solve the nonlinear and non-convex optimization problem of (1) for estimating the optical flow at the superpixel level. Starting from the warping-based optical flow estimation [3] as initialization, we apply the technique of gradient descend to solve $\mathbf{u_i}$ for the ith superpixel:

$$\mathbf{u}_i^{t+1} = \mathbf{u}_i^t - \lambda(\frac{\partial E(\mathbf{u}_i)}{\partial \mathbf{u}_i}) = \mathbf{u}_i^t + \lambda'(\frac{\sum\limits_{p\in B_i^C} \mathbf{x}_p^i \psi_p^{i^C} g(\left\|\mathbf{x}_p^i\right\|_2)}{\sum\limits_{p\in B_i^C} \psi_p^{i^C} g(\left\|\mathbf{x}_p^i\right\|_2)} + \sum_{j\in N_i^G} \frac{w'^G_{(i,j)} \sum\limits_{p\in B^G_{(i,j)}} \mathbf{x}_p^{(i,j)} \psi_p^{(i,j)^G} g(\left\|\mathbf{x}_p^{(i,j)}\right\|_2)}{\sum\limits_{p\in B^G_{(i,j)}} \psi_p^{(i,j)^G} g(\left\|\mathbf{x}_p^{(i,j)}\right\|_2)} + \sum_{j\in N_i^S} w'^S_{(i,j)}(\mathbf{u}_j^t - \mathbf{u}_i^t)), \quad (5)$$

where λ is the step size for the optimization. Variables $\mathbf{x}_p^i = \mathbf{x}_p - (\mathbf{x}_i + \mathbf{u}_i^t)$ and $\mathbf{x}_p^{(i,j)} = \mathbf{x}_p - (\mathbf{x}_{(i,j)} + \mathbf{u}_i^t)$ are the distances between the pixel p and the corresponding estimated superpixel or boundary centers, respectively. Functions $\psi_p^{i^C} = \exp\left(-\frac{\left\|\mathbf{I}_2(\mathbf{x}_p) - \mu_i^C\right\|_2}{\beta^C}\right)$, and $\psi_p^{(i,j)^G} = \exp\left(-\frac{\left\|\nabla\mathbf{I}_2(\mathbf{x}_p) - \mu^G_{(i,j)}\right\|_2}{\beta^G}\right)$ calculate the associated color and gradient affinity weights, respectively. Parameters $\lambda', w'^G_{(i,j)}$ and $w'^S_{(i,j)}$ are the weights associated and proportional to $\lambda, w^G_{(i,j)}$ and $w^S_{(i,j)}$.

From (5), it can be seen that we approach the optical flow estimation problem as solving the task of mean-shift tracking [12] at the superpixel level using multiple visual features. With the introduced smoothness constraint, we alleviate the problem of unsuccessful matching/tracking across videos due to insufficient resolution caused by noise or motion blur. In addition, since our search range (i.e., B_i^C) is based on the size of superpixels and the associated color difference, we alleviate local minimum problems (which cannot be easily addressed using standard mean-shift algorithms).

When applying (5) to update the estimated superpixel-based optical flow, the data and smoothness terms in (1) are jointly optimized. As a result, we are able to preserve the smoothness of the calculated optical flow during the feature matching process. Moreover, like particle swarm optimization [13], this optimization process tends to search for the *global* minimum of (1) by passing the matching information of superpixels which are farther away but with similar color information. This is another reason why improved estimation results can be expected.

We note that, our optimization process would terminate if $\mathbf{u}_i$ converges or a maximum number of iterations (we set 400 in our experiments) is reached. The estimated superpixel-based optical flow will be refined by the technique of Classic-C-brightness (see [14] for details) for reducing the blocking artifacts due to image over-segmentation.

2.3. Occlusion Handling in Optical Flow Estimation

Typically, videos with large displacement would be accompanied with severe occlusion effects, which poses a challenging task for optical flow estimation. In LDOF [5], consistency check in a backward direction during descriptor matching has been successfully applied to alleviate the above problem. However, this technique cannot be easily extended to videos with insufficient resolution. In our work, we perform forward and backward optical flow estimation to identify optical flow from occluded regions. This is due to the fact that optical flow estimation for an occluded region would only fail in one of the directions. As suggested by [15], a probability function can be calculated by the above results (the probability function output would indicate how likely the associated superpixel is not occluded). Thus, to handle possible occlusion and produce improved estimation results, we multiply the data terms in (1) by the above probability values before solving the optimization problem.

3. EXPERIMENTS

To evaluate the performance of our method, we consider a 3D animation video MPI-Sintel dataset which is recently released by [17]. Although the videos in this dataset are of size 1024 $\times$ 436 pixels per frame, foreground or background regions in some videos are typically observed to exhibit significant motion. As a result, effects like low color contrast, large displacement, articulated motion patterns, or motion blur make the optical flow estimation very difficult. Since the ground truth optical flow information is available, we are able to perform qualitative and quantitative comparisons with other state-of-the-art methods. When evaluating our approach, we select and fix the parameters like weights for energy terms for producing the best estimation results for all videos [1].

We select seven video sequences from this dataset: *ambush 2, ambush 5, ambush 6, cave 2, cave 4, market 5*, and *market 6*. For videos *cave4* and *market6*, large motion information was only observed for small foreground regions. On the other hand, videos *ambush2* and *ambush5* exhibit large displacement motion from background regions with low color contrast or heavy occlusion. In our experiments, we select the

[1] Code available at: http://mml.citi.sinica.edu.tw/#tabs_project

Table 1. Comparisons of end point error (EPE) for different methods. Note that * indicates the results of LDOF on manually-blurred videos, and the numbers in bold denote the best results for the corresponding videos.

Method \ Sequence	ambush2	ambush5	ambush6	cave2	cave4	market5	market6	Avg
Warping [3]	**65.45**	42.13	59.62	73.51	24.11	54.24	23.68	48.96
Horn+Schunck [16]	68.95	46.51	58.70	70.36	22.27	48.63	24.11	48.51
Classic+NL-fast [14]	69.02	45.61	57.31	70.21	21.06	51.05	24.48	48.39
LDOF [5]	73.24	49.17	60.71	72.03	29.66	53.80	33.02	53.09
LDOF* [5]	68.95	37.75	56.60	58.06	**19.03**	43.14	22.74	43.75
Ours	66.80	**35.63**	**45.09**	**45.63**	19.87	**33.34**	**19.34**	**37.96**

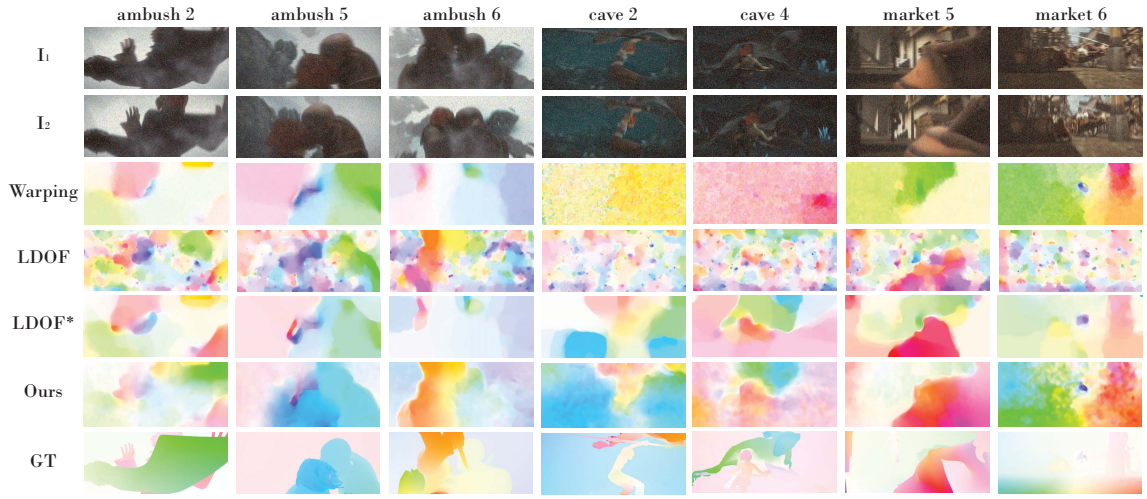

Fig. 2. Example optical flow estimation results for different approaches. Note that GT denotes the ground truth optical flow.

top 20% video frames from each video based on the variations of the ground truth optical flow (without those in which the foreground object moves out of the frame). To deal with low color contrast, motion blur, etc. noise effects, we apply motion blur kernel in both directions with window sizes as 2% of the image width, and add salt & pepper noise with density 0.2 for degrading the quality of videos.

We compare our results with those produced by different optical flow methods: Warping [3], Horn+Schunck [16], Classic+NL-fast [14], and LDOF [5]. We do not require the prior knowledge on the types of noise (like motion blur or other artifacts) as [18] and [19] did. Since LDOF is based on descriptor matching, its performance will be sensitive to descriptor extraction (especially for noisy videos). In order to provide additional robustness for LDOF, we also perform LDOF on blurred videos using Gaussian kernels of $\sigma = 10$ (denoted as * in Table 1 and Figure 2), which is expected to outperform the standard LDOF on raw noisy videos.

To quantitatively compare the estimation performance, Table 1 lists the average values of end point error (EPE), which indicates the difference (distance) between the ground truth optical flow and the estimated one. From this table, it is clear that our approach achieved improved or comparable performances, while descriptor-matching based methods generally produced the poorest results. Since we do not require such matching techniques or the prior knowledge of noisy type in videos, significant improvements can be obtained especially for videos with large displacements (i.e., *ambush6, cave2*, and *market5*). This confirmed that our approach is able to alleviate local minimum problems caused by linear approximation of traditional optical flow algorithms. As example results shown in Figure 2, we see that the estimated optical flow produced by our method was very similar to the ground truth, while most image details were well preserved. From the above experiments, the effectiveness and robustness of our approach can be successfully verified.

4. CONCLUSION

We presented a superpixel-based optical flow estimation algorithm particularly for videos with large displacement and insufficient resolution. Our proposed algorithm solves an energy minimization problem, which jointly optimizes data and smoothness terms using color and gradient features at the superpixel level. As an extension of the mean-shift algorithm, our approach expanded the range for optical flow estimation and alleviated potential local minimum problems. Experiments on the MPI-Sintel dataset confirmed that our method quantitatively and qualitatively outperformed coarse-to-fine warping or descriptor matching based approaches.

Acknowledgement This work is supported in part by National Science Council of Taiwan via NSC100-2221-E-001-018-MY2.

5. REFERENCES

[1] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, 1981.

[2] L. Alvarez, J. Weickert, and J. Snchez, "Reliable estimation of dense optical flow fields with large displacements," *IJCV*, 2000.

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, 2004.

[4] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell*, 2011.

[5] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011.

[6] L. Xu, J. Chen, and J. Jia, "A segmentation based variational model for accurate optical flow estimation," in *ECCV*, 2008.

[7] C. Lei and Y.-H. Yang, "Optical flow estimation on coarse-to-fine region-trees using discrete optimization," in *ICCV*, 2009.

[8] F. Steinbrücker, T. Pock, and D. Cremers, "Large displacement optical flow computation without warping," in *ICCV*, 2009.

[9] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.

[10] W.-T. Li, H.-T. Chang, H. Lyu, and Y.-C. F. Wang, "Automatic saliency inspired foreground object extraction from videos," in *ICIP*, 2012.

[11] P. Krhenbhl and V. Koltun, "Efficient nonlocal regularization for optical flow," in *ECCV*, 2012.

[12] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking.," 2003.

[13] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *ICNN*, 1995.

[14] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *CVPR*, 2010.

[15] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[16] D. Sun, S. Roth, J. P. Lewis, and M. J. Black, "Learning optical flow.," in *ECCV*, 2008.

[17] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012.

[18] T. Portz, L. Zhang, and H. Jiang, "Optical flow in the presence of spatially-varying motion blur.," in *CVPR*, 2012.

[19] H. Scharr and H. Spies, "Accurate optical flow in noisy image sequences using flow adapted anisotropic diffusion.," 2005.