Foreground object extraction through motion segmentation

Yinhui Zhang (张印辉) and Zifen He (何自芬)*

Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China

**Corresponding author: zyhhzf1998@163.com*

Received April 11, 2014; accepted July 16, 2014; posted online January 16, 2015

We present a method to extract foreground object regions efficiently from image sequences. Scale-invariant feature transform algorithm is adopted to estimate the descriptor firstly by matching between two consecutive frames. Given local descriptor matching results, dense motion vector of each pixel is calculated by large displacement optical flow with variational optimization, which integrates detailed descriptors into the variational model. Then the foreground object boundaries and regions are detected by computing the optical flow gradient and magnitude. Experiments demonstrate that the method can achieve better segmentation results than alternative methods and adapts well to moving objects in relatively stationary background image sequences.

OCIS codes: 100.2000, 100.2960.
doi: 10.3788/COL201513.S11002.

Image segmentation is one of the most fundamental problems in computer vision society. The goal of segmentation is to find distinct homogeneous regions in the image field. Recently, image segmentation has been used in object recognition[1], activity analysis[2] as well as in stereo vision[3].

Segmentation by utilizing motion cues is an important branch in image segmentation as visual motion tends to be a powerful cue for low-level segmentation in human visual system. Optical flow, as a key motion cue, by definition, is the pattern of apparent motion of objects in a visual scene caused by relative motion between an observer and the scene. Optical flow is derived from a sequence of images that represent the motion, position, and orientation of an object. Its capability has been demonstrated in human visual system in various psychological experiments[4]. For instance, when foreground object is stationary relative to background, the response of motion is approximately zero and motion boundary is invisible to the observer. In contrast, when an object is moved relative to each other the object boundary becomes obvious immediately to the observer. This is very important for practical systems in industry. For this reason, optical flow-based motion segmentation was used in robotic vision fields such as object detection and visual servo, image domain plane extraction, robot navigation, and visual odometry.

The importance of optical flow as a cue for segmentation is widely accepted by researchers. Optical flow was introduced by American psychologist James J. Bibson in the 1940s to describe visual stimulus provided to animals moving throughout the world. Recently, the popular approach to estimate optical flow is based on variational methods proposed by the seminal work in Ref. [5]. Consequently, this method is extended and improved to overcome its limitations. By imposing nonquadratic penalty into pairwise and data energy terms motion discontinuities can be compensated. Then nonlinear constancy constraints are incorporated to deal with large displacements. However, a key problem in variational optical flow is the local optimization problem. That is, the optimization result is sensitive to initialization. The algorithm usually selects the local minima with the smallest motion.

By guessing initialization of motion at coarse scale, the coarse-to-fine warping[6] method can alleviate local minima problem. Firstly, this method estimates the initialization by utilizing motion of large structures in images. Secondly, motion field is successively refined by incorporating new evidences of smaller parts of objects. When smaller parts move approximately same as the main part of foreground structures, the coarse-to-fine warping method works well. However, this method cannot cope with the problem when smaller parts move relatively larger than main structure. In this circumstance, traditional coarse-to-fine warping methods always do not work in practice. Small parts of object can move very fast with respect to the main structures, violating the assumption that the main structure should be a good initialization of the whole motion. Most recently, video object segmentation is performed in the object proposal domain. For example, a layered directed acyclic graph (DAG)-based object segmentation has been introduced in Ref. [7] to detect and segment the salient objects in image sequences. But the object ranking problem is also ill-posed in that the hypotheses ranking is hard to compute in a principled way.

We use the method proposed by Brox *et al.*[8] to deal with motion detection with large displacement optical flow. The key idea is to employ descriptor instead of single pixel matching to ensure global matching. Popular local descriptors such as scale-invariant feature transform (SIFT) and histogram of oriented gradient are usually good enough for global matching. In this

letter, we employ SIFT descriptor matching method to obtain pixel correspondences between two frames. This can be explained by the successful descriptor-based methods in the field of image analysis.

The key problems in adapting descriptor matching to optical flow-based image segmentation lie in several aspects. One of the difficulties in local descriptor matching problem is that some of the counterparts of descriptors are missing due to image occlusions. By integrating the correspondences from local descriptor matching into a variational optical flow framework, the method in Ref. [8] can benefit both from large displacement descriptor matching and from accurate variational techniques.

Flow chart of the proposed method is shown in Fig. 1. We first introduce SIFT algorithm to estimate the descriptor matching. Then calculate dense motion vector of each pixel through large displacement optical flow through variational optimization, which integrates detailed descriptors into the variational model. Then we introduce the foreground object boundary detection from the gradient and magnitude of the optical flow vectors.

Earlier works use corner detector to search interest points for image matching. In practice, the Harris corner detector is sensitive to scale changes, thus it is not applicable to image matching between images with different sizes. The SIFT detector is invariant to image scale and rotation. Moreover, it facilitates robust matching across affine distortion, viewpoint change, and illumination change. We employ the algorithm proposed in Ref. [8] to estimate the local descriptor of detected key points. The scale space is given as

$$L(x,y,\sigma) = G(x,y,\sigma)*I(x,y),$$

where $G(x,y,\sigma)$ is variable-scale Gaussian and $I(x,y)$ is the input image.

To detect key points in the image, scale-space extrema is used by the author on the convolution result between the original image and the difference-of-Gaussian function: $G(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y)$. The scale-space extrema is efficient to compute and the difference-of-Gaussian is an approximation to the Laplacian of Gaussian function. Once the difference-of-Gaussian is computed, each sample point is compared with its 26 neighbors both in the same image scale and in the neighboring scales. By this way, the local maxima and minima of the difference-of-Gaussian is obtained.

Orientation histograms are used to represent the local descriptor. The image gradient magnitudes and orientations are sampled around the key point. A Gaussian weighting parameter is used to assign a weight to the magnitude of each sample point for the purpose of avoiding non-smooth changes in the descriptor. In our experiments we used 4×4 array of histograms with eight orientation bins. Therefore, each feature vector of a key point has 4×4×8 = 128 elements.

We introduce the main optical flow algorithm proposed by Lowe[9]. Let I_1, I_2 be two consecutive frames in the image sequence. $\mathbf{x}=(x,y)^T$ denotes a pixel location in the image domain and $\mathbf{w}=(u,v)^T$ denotes the optical flow vector field to be estimated. The pixels in the two frames correspond to optical flow vector. One assumption is that the corresponding points should have similar color or gray values. To integrate this constraint, the color data term E_{color} is incorporated by $\int_\Omega \Psi \mid I_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) - I_1(\mathbf{x}) \mid \mathrm{d}\mathbf{x}$. Similarly, gradient deviation of corresponding points E_{grad} is integrated into the energy function to deal with illumination effects. Precisely, the gradient data term is given as $\int_\Omega \Psi \mid \nabla I_2(\mathbf{x}+\mathbf{w}(\mathbf{x})) - \nabla I_1(\mathbf{x}) \mid^2 \mathrm{d}\mathbf{x}$. Enforcing only the color and gradient constraints usually leads to ambiguous optical flow field vectors because they do not take into account the similarity information between neighboring pixels. To this end, the regularity constraints E_{smooth} are added to the energy function by penalizing the total variation of the flow field, which is defined as

$$\int_\Omega \Psi(\mid \nabla u(\mathbf{x}) \mid^2 + \mid \nabla v(\mathbf{x}) \mid^2)\mathrm{d}\mathbf{x}.$$

The result of point correspondence through local descriptor matching is integrated into the variational optical flow energy function by

$$E_{\text{match}}(\mathbf{w}) = \int_\Omega \delta(\mathbf{x})\rho(\mathbf{x})\Psi(\mid \mathbf{w}(\mathbf{x}) - \mathbf{w}_1(\mathbf{x}) \mid^2)\mathrm{d}\mathbf{x}, \quad (1)$$

where $\mathbf{w}_1(\mathbf{x})$ denotes the pixel correspondence vectors result from SIFT descriptor matching at location $\mathbf{x}$. $\delta(\mathbf{x})$ is an indicator function to indicate the presence of a SIFT descriptor at pixel $\mathbf{x}$. $\rho(\mathbf{x})$ denotes the matching score at pixel $\mathbf{x}$. The last term is used to penalize deviation of descriptor feature vectors that are linked by SIFT descriptor matching. In particular, this term is given as

$$E_{\text{desc}}(\mathbf{w}_1) = \int_\Omega \delta(\mathbf{x})(\mid \mathbf{f}_2(\mathbf{x}+\mathbf{w}_1(\mathbf{x}) - \mathbf{f}_1(\mathbf{x}) \mid^2)\mathrm{d}\mathbf{x}, \quad (2)$$

where $\mathbf{f}_1(\mathbf{x})$ and $\mathbf{f}_2(\mathbf{x})$ denote the feature vector of SIFT descriptors in frame 1 and frame 2 at pixel $\mathbf{x}$. Adding

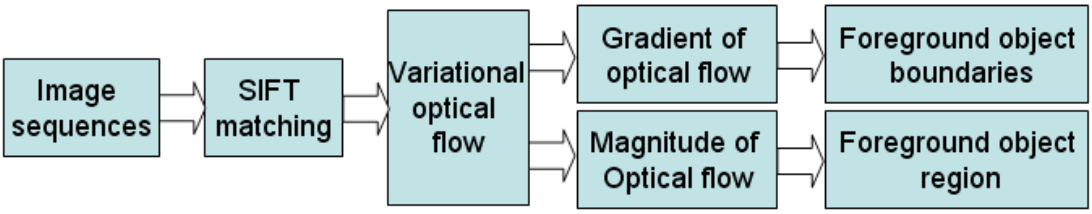

Fig. 1. Flow chart of the proposed motion segmentation method.

the five energy terms together constitute the variational optical flow energy function that integrates not only pixels' smoothness but also higher order descriptor matching constraints.

Given the optical flow field computed under motion and descriptor matching conditions, we employ a standard motion estimation method proposed in Ref. [10], that is, Foreground boundaries are estimated by computing the magnitude of the gradient of the optical flow field by

$$b(x,y) = 1 - \exp(-\lambda \,||\, \nabla \mathbf{w}(x,y) \,||), \tag{3}$$

where $b(x,y) \in [0,1]$ is the intensity of the motion boundary at pixel location (x,y). The intensity of the motion is particularly adapted to object boundaries with high motion speeds. However, for intermediated intensity values of the motion boundary, it is difficult to discriminate whether it belongs to object boundaries. To this end, the authors in Ref. [10] also proposed another criterion which is defined as

$$b'(x,y) = 1 - \exp(-\lambda' \max_{y \in N} \theta^2(x,y)), \tag{4}$$

where $\theta(x,y)$ denotes the angle between two flow vectors located at (x,y). The intuitive of this criterion is that if a point is moving rapidly than all its neighbors, then it is likely to be a motion boundary. Using this criterion, the boundary detection method enables to detect boundaries even if the foreground is not moving rapidly.

However, as also indicated in Ref. [10], the motion boundaries tend to produce false positives in static image regions due to the noisy optical flow field estimated using the variational optimization method. For this reason, we further propose motion region criterion to provide a more robust foreground extraction method. We define the motion region as

$$r(x,y) = || \mathbf{w}(x,y) ||_2 > \tau, \tag{5}$$

where $||\cdot||_2$ denotes the L2-norm and τ is the threshold selected to obtain motion regions. Intuitively, larger threshold τ results motion regions with higher moving speeds and vice versa.

We demonstrated the performance using the object boundary detection method on the dog sequence. A sample of the images in the dog sequences is shown in Fig. 2. In this sequence, we can observe that the object has similar appearance with the background and segmentation, and using only color cue has always failed to segment the dog from the road background. The experiments were conducted on an Intel Core i5 (2.67 GHz) machine with 3 GB memory using MATLAB.

In our experiments, we first detect the descriptor matching of two consecutive frames using SIFT algorithm. The detected key points and their respective descriptors of the first two frames are shown in Figs. 3(a) and (b). In general, 319 key points are detected in frame #1 and 345 key points are detected in frame #2. Each key point is represented by a 4×1 vector which contains the x- and y-locations, the scale as well as the orientation of the detected key point. For instance, the first key point in frame #1 is located at (5, 13) at scale of 2.0993 with orientation of 0.8877 rad. In addition, histograms are computed from magnitude and orientation values of samples in a 16×16 region around the key point such that each histogram contains samples from a 4×4 subregion of the original neighborhood region. Consequently, two sets of key points are matched in accordance with the L2-distance between their respective descriptors. The matching results are illustrated in Figs. 3(c) and (d), in which 58 pairs of key points are estimated. We observe that most of the key points are correctly matched.

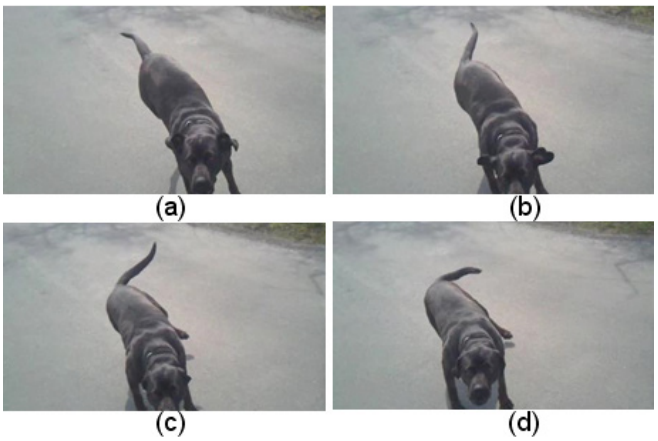

Fig. 2. Sample of original images in the dog video sequence of (a) frame #1, (b) frame #2, (c) frame #3, and (d) frame #4.

Given the SIFT matching results, we can set up the parameters in the energy function, in particular, pixel

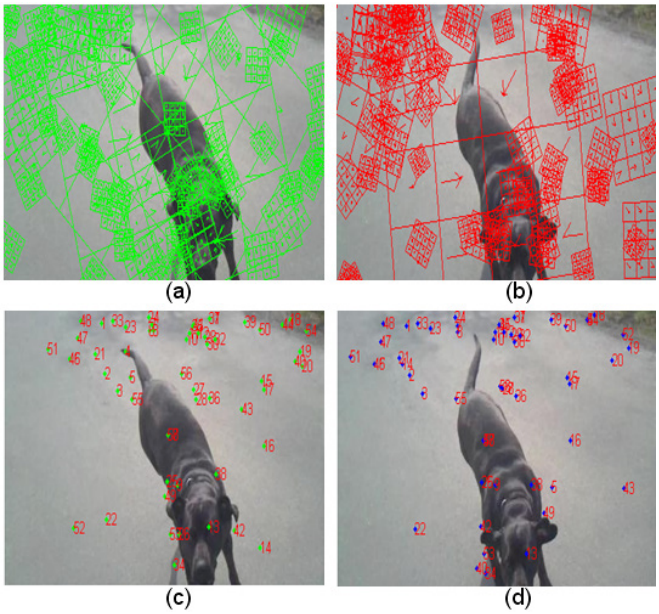

Fig. 3. SIFT descriptor matching process: (a) and (b) are SIFT descriptors overplot on the first two frames, and (c) and (d) are SIFT descriptor matching results.

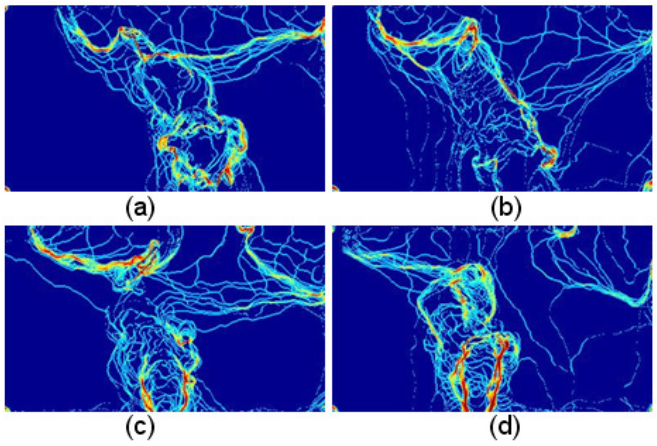

Fig. 4. Motion boundary by computing gradient magnitude of the optical flow field of (a) frame #1, (b) frame #2, (c) frame #3, and (d) frame #4.

correspondence vector $\mathbf{w}_1(\mathbf{x})$, indicator function $\delta(\mathbf{x})$, and the matching score $\rho(\mathbf{x})$ at each pixel $\mathbf{x}$ as well as the feature vector of SIFT descriptors in two frames $\mathbf{f}_1(\mathbf{x})$ and $\mathbf{f}_2(\mathbf{x})$. Once the parameters are set up, we then minimize the sum of the five energy terms: E_{color}, E_{grad}, E_{smooth}, E_{desc}, and E_{match} by variational minimization algorithm. The solution of the variational minimization is the optimal optical flow vector field $\mathbf{w}^*=(u,v)^T$.

At the motion boundary detection stage, we first choose the parameters in motion boundary detection Eqs. (1) and (2). The two parameters λ and λ' control the declining speed of the value of boundary mapping to the interval [0, 1]. In particular, we find $\lambda=0.5$ and $\lambda'=0.7$ will always achieve good boundary detection results.

Figure 4 shows the detected object boundary $b(x,y)$ in Eq. (3) and the intensity of the motion boundary at pixel location (x,y). From the motion boundaries we can observe that the intensity of the motion boundary is particularly adapted to object boundaries, which have high motion speeds. Given these motion boundaries,

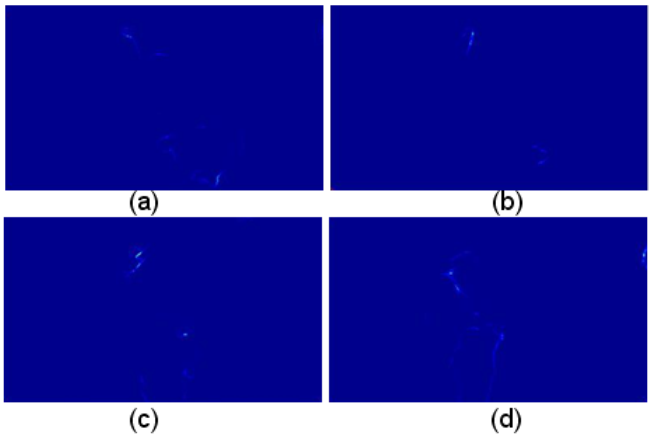

Fig. 5. Motion rotation boundary $b'(x,y)$ of (a) frame #1, (b) frame #2, (c) frame #3, and (d) frame #4.

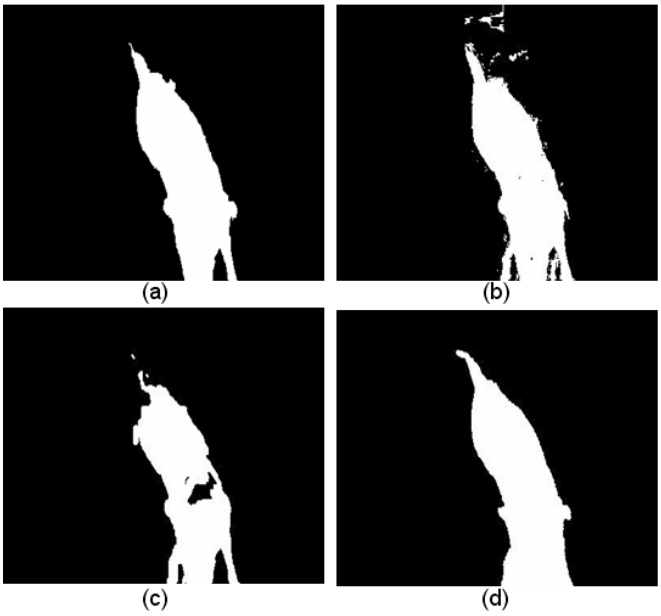

Fig. 6. Foreground segmentation results of frame #1 using different methods: (a) ground truth, (b) GMM results, (c) DAG[7] results, and (d) our segmentation results.

we obtain good foreground object boundaries cues that help future object segmentation or object tracking.

Figure 5 shows the detected object boundary $b'(x,y)$ in Eq. (4) and the intensity of the motion boundary at each pixel location (x,y). From the motion boundaries $b'(x,y)$ we can observe that the high intensity appears at the locations where a point is moving rapidly than all its neighbors, for example, the higher values at the head and tail of the foreground object dog are evident. Using this criterion, the boundary detection method enables to detect boundaries even if the foreground is not moving rapidly.

In this dog video sequence, the motion region criterion does not provide an ideal motion segmentation result that is as obvious as both motion boundary $b(x,y)$ and motion rotation boundary $b'(x,y)$. The reason might lies in the non-homogeneous motion of the foreground object so as to a single threshold global threshold leads to over-segmentation of the motion vector field.

We compare segmentation results of the proposed method with that of Gaussian mixture model (GMM) as well as DAG[7]. In all experiments reported here, the threshold τ is set to 0.5. The segmentation results are shown in Fig. 6. Qualitatively, the proposed method achieves relatively good segmentation results than GMM and DAG. We further compute false error rate (FER) of segmentation results of the first five frames in Table 1. The criterion of FER is consistent with that of the qualitative segmentation results. The proposed method obtains the smallest FER than both GMM and DAG methods over five frames, which demonstrates the effectiveness of the proposed motion segmentation method.

Table 1. FER of Segmentation Results of the First Five Frames

Frame	**#1**	**#2**	**#3**	**#4**	**#5**
GMM	2126	1400	1481	2074	3004
DAG	2015	1264	1376	1671	1970
Ours	1286	1067	676	1340	370

In conclusion, we use SIFT algorithm to estimate the descriptor matching between two consecutive frames. We then calculate the dense motion vector of each pixel through large displacement optical flow through variational optimization. The foreground object boundaries are detected by computing the gradient of the optical flow vectors. Thus, motion boundary, motion rotation boundary as well as motion region can be achieved in the selected image sequence. Experiments demonstrate that the method can achieve both motion boundaries and the region that adapts well to moving objects in the relatively stationary background image sequences.

This work was supported by the National Science Foundation of China under Grants Nos. 61461022 and 61302173. Yinhui Zhang was supported by the China Scholarship Council (No. 201208535035) while studying as a visiting scholar at University of Miami.

References

1. J. Winn and N. Jojic, in *Proceedings of IEEE International Conference on Computer Vision* 756 (2005).
2. C. Stauffer and W. E. L. Grimson, IEEE Trans. Pattern Anal. Mach. Intell. **22,** 747 (2000).
3. J. A. Delmerico, P. David, and J. J. Corso, Image Vis. Comput. **31,** 841 (2013).
4. D. W. Murray and B. F. Buxton, Trans. Pattern Anal. Mach. Intell. **9,** 220 (1987).
5. B. Horn and B. Schunck, Artif. Intell. **17,** 185 (1981).
6. T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, in *Proceedings of Eighth European Conference On Computer Vision* 25 (2004).
7. D. Zhang, O. Javed, and M. Shah, in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition* 628 (2013).
8. T. Brox and J. Malik, Trans. Pattern Anal. Mach. Intell. **33,** 500 (2011).
9. D. G. Lowe, Int. J. Comput. Vis. **60,** 91 (2004).
10. A. Papazoglou and V. Ferrari, in *Proceedings of IEEE International Conference on Computer Vision* 1777 (2013).