Generalized Background Subtraction Using Superpixels with Label Integrated Motion Estimation

Jongwoo Lim[1] and Bohyung Han[2]

[1] Division of Computer Science and Engineering, Hanyang University, Seoul, Korea
[2] Department of Computer Science and Engineering, POSTECH, Korea
jlim@hanyang.ac.kr, bhhan@postech.ac.kr

Abstract. We propose an online background subtraction algorithm with superpixel-based density estimation for videos captured by moving camera. Our algorithm maintains appearance and motion models of foreground and background for each superpixel, computes foreground and background likelihoods for each pixel based on the models, and determines pixelwise labels using binary belief propagation. The estimated labels trigger the update of appearance and motion models, and the above steps are performed iteratively in each frame. After convergence, appearance models are propagated through a sequential Bayesian filtering, where predictions rely on motion fields of both labels whose computation exploits the segmentation mask. Superpixel-based modeling and label integrated motion estimation make propagated appearance models more accurate compared to existing methods since the models are constructed on visually coherent regions and the quality of estimated motion is improved by avoiding motion smoothing across regions with different labels. We evaluate our algorithm with challenging video sequences and present significant performance improvement over the state-of-the-art techniques quantitatively and qualitatively.

Keywords: generalized background subtraction, superpixel segmentation, density propagation, layered optical flow estimation.

1 Introduction

Moving object detection in videos is a critical step to many computer vision problems such as visual tracking, scene understanding, human motion analysis, unmanned vehicle navigation, event detection and so on. One of the approaches for this task is background subtraction, also known as foreground/background segmentation, which is typically based on appearance modeling and update of foreground and background in local or global areas. Traditionally, background subtraction has been investigated in a stationary camera environment [1–6], but researchers recently started to study the problem with a moving camera. Background subtraction with a freely moving camera is obviously more challenging particularly due to unreliable motion estimation caused by fast motion,

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 173–187, 2014.

occlusion, motion blur, etc. Consequently, a simple extension of background subtraction algorithms to a moving camera environment would fail easily because appearance models are prone to be contaminated by inaccurate image registration across frames. Our goal in this work is to tackle the more challenging foreground/background segmentation problem, where we propose a superpixel-based modeling of appearance and motion and a separate foreground/background motion estimation using segmentation mask.

There are several closely related works for background subtraction in videos captured by a moving camera. The most primitive algorithms are probably motion segmentation and its extensions [7–9], and they separate foreground from background based on homography or homography+parallax. However, they assume that the dominant motion is from background and only residual motions belong to foreground objects, which may not be true in practice. A few approaches to combine motion estimation and appearance modeling are recently proposed for online background subtraction [10, 11], but they rely on robust estimation of long term trajectories such as particle video [12]. In [13, 14], block-based density propagation techniques are proposed for generalized background subtraction[1], but their algorithms are complex and involve many free parameters; the performance in a general setting may not be consistent. On the other hand, [15] employs a matrix factorization technique with low rank and group sparsity constraints of long-term trajectories, and [16] proposes a multi-layer segmentation algorithm by label propagation from given sparse trajectories. However, both methods run offline and rely on robust trajectory estimation. None of previous works investigate the interaction between segmentation and motion estimation although they are tightly coupled since the quality of estimated motion can be improved by human annotated foreground and background boundaries as discussed in [17].

Contrary to existing techniques, our generalized background subtraction algorithm utilizes the segmentation mask to compute foreground and background motion fields and use them to maintain more accurate appearance and motion models. Given pixelwise motion vectors, the proposed algorithm propagates foreground and background appearance models of the previous frame, which are defined in each superpixel, through a sequential Bayesian filtering. It also builds motion models for each superpixel based on the motion vector observations. Pixelwise foreground/background likelihood is computed based on the appearance and motion models, and segmentation labels are determined by binary belief propagation. Label estimation in each pixel triggers update of motion and appearance models in each superpixel, and the final label of the frame is obtained after convergence via a few iterations.

The overview of our algorithm is illustrated in Figure 1. Our algorithm is differentiated from previous works such as [13, 14] in the sense that there is interaction between foreground/background segmentation and motion estimation, and appearance and motion modeling is performed on homogeneous regions

[1] This term means background subtraction in a moving camera environment, and is first used in [13].

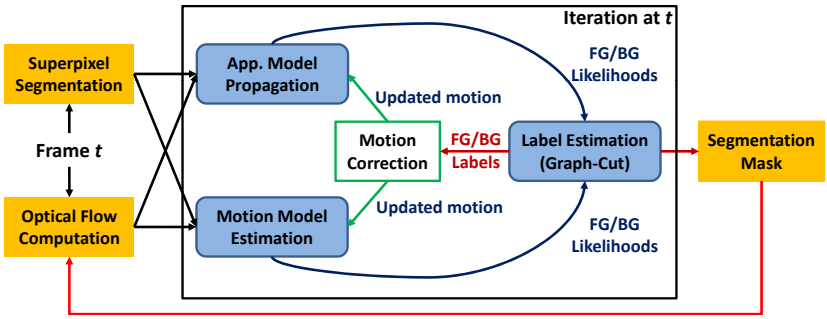

Fig. 1. Overview of our algorithm. Main contributions in our algorithm are highlighted with yellow boxes and a red line. Note that we compute separate foreground and background motion fields.

(superpixels) for more efficient and accurate estimation. The advantages and characteristics of the proposed algorithm are summarized below:

- Our algorithm employs segmentation mask to estimate foreground and background motion fields separately and avoids motion smoothing across regions with different labels. It improves segmentation quality by maintaining more accurate appearance and motion models.
- Instead of a regular grid-based modeling as in [13, 14], superpixel-based modeling is employed for reliable density estimation; the observations of color and motion in a superpixel are coherent, and simple density representations are sufficient for accurate modeling.
- The proposed algorithm is more efficient than [13] by using simple histograms in density representation and avoiding complex inference procedures in density propagation. The performance is improved significantly with fewer number of free parameters.

The organization of this paper is as follows. Section 2 summarizes superpixel segmentation method used in our algorithm and Section 3 describes our motion estimation technique based on segmentation mask. The main background subtraction algorithm is presented in Section 4, where model construction, density propagation, and label estimation with likelihood computation are discussed in detail. Section 5 illustrates experimental results with real videos.

2 Superpixel Segmentation

Contrary to [13, 14], where the frames are divided into regular rectangular grid blocks, we employ a superpixel segmentation for modeling appearance and motion. In a superpixel, appearance and motion are likely to be homogeneous hence estimated density functions are to be more reliable and accurate given a limited number of pixels.

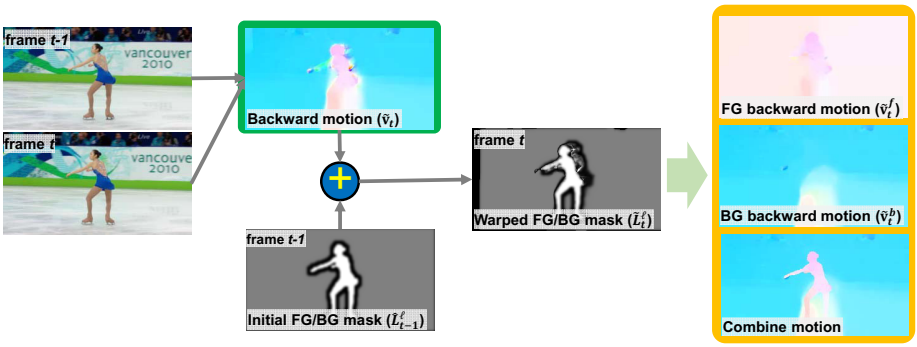

Fig. 2. Separate foreground and background motion estimation by segment mask propagation from the previous label. Note that [13, 14] use backward motion in green plate, but our algorithm employs a separate motion field for each label in orange plate. The combined motion shows clearer motion boundary. The motion images are color-coded to visualize the direction and magnitude, and the white, gray, and black areas in the mask images represent the foreground, background, and ambiguous regions respectively.

We use the ERS superpixel segmentation algorithm [18], due to its simplicity and perceptually good performance, which formulates the superpixel segmentation as a graph partitioning problem. Given a graph $G = (V, E)$ and the number of superpixels K, the goal is to find $A \subseteq E$ such that the resulting graph $\widetilde{G} = (V, A)$ contains K connected subgraphs. Note that a vertex corresponds to a pixel in image and edges are typically constructed by the 4-neighborhood system, where the weight of an edge is computed by the similarity between the features observed at the connected vertices. The objective function to solve the graph partitioning problem is given by

$$\begin{aligned} \max_{A} \quad & \mathcal{H}(A) + \lambda \mathcal{B}(A) \\ \text{s.t} \quad & A \subseteq E \text{ and } N_A \geq K \end{aligned} \tag{1}$$

where $\mathcal{H}$ and $\mathcal{B}$ denote entropy rate of random walk and balancing term respectively, and N_A is the number of connected components in $\widetilde{G}$. The entropy rate term encourages compact and homogeneous segments and the balancing function is to segment with similar sizes.

Although the exact optimization is difficult, it can be solved by a greedy algorithm efficiently and the solution by this approach always provides 0.5 approximation bound. We refer the reader to [18] for more details.

3 Motion Estimation with Segmentation Mask

Unlike existing approaches that use a single motion field, we propose to estimate two separate motion fields for foreground and background of the scene. Due to

the smoothness assumption that most optical flow algorithms adopt, the motion field near motion boundary or depth discontinuity tend to be over-smoothed and blurred. For foreground/background segmentation, motion boundaries are the most important regions and inaccurate motions near the area often produce incorrect labels. It is because the erroneous motion vectors compromise the estimated motion models and corrupt the propagated appearance models.

Our idea is motivated by [17], where the accuracy of motion estimation is improved significantly with object boundary annotation by human. We compute separate motion fields for foreground and background using the corresponding segmentation masks estimated by our algorithm in the previous frame. To compute the backward motion field $\mathbf{v}_t^\ell$ for label $\ell \in \{f, b\}$ at frame t, we need the warped foreground/background segmentation mask $\tilde{L}_t^\ell$, which is estimated from the mask in the previous frame L_{t-1}^ℓ and the backward motion without segmentation mask $\mathbf{v}_t$ using [19], as

$$\tilde{L}_t^\ell(\mathbf{x}) = \hat{L}_{t-1}^\ell(\mathbf{x} + \mathbf{v}_t(\mathbf{x})), \tag{2}$$

where $\hat{L}_{t-1}^f = L_{t-1}^f$ and $\hat{L}_{t-1}^b = \xi(L_{t-1}^b, r)$. The morphological erosion function with label L and radius r denoted by $\xi(L, r)$ is performed to reduce the effect of occluded or uncovered background pixels.

Once the segmentation mask $\tilde{L}_t^\ell$ is given, we compute the backward motion field $\mathbf{v}_t^\ell$ for each foreground and background label using [19], which ignores the observations in the unset region and propagates motions spatially from the neighboring pixels. Figure 2 illustrates the procedure to compute backward foreground/background motion fields.

4 Background Subtraction Algorithm

Our foreground/background segmentation algorithm is composed of the following three steps: model construction, density propagation, and label estimation with likelihood computation. These procedures are repeated in each frame until convergence. We describe technical details of each of these three steps.

4.1 Appearance and Motion Models

We construct appearance and motion models for foreground and background in each superpixel. These models are two main factors to determine the label of each pixel, and accurate and efficient estimation of the models is crucial for the success of our algorithm. We employ simple histogram to represent appearance and motion models since basic operations on histogram(s) such as addition, product, and convolution can be implemented straightforwardly and performed with low computational cost. In addition, histogram is advantageous compared to continuous distributions such as kernel density estimation used in [13] especially when feature dimensionality is low.

Suppose that we have maintained the posterior of appearance corresponding to the ith superpixel at frame t, which is a normalized histogram denoted by

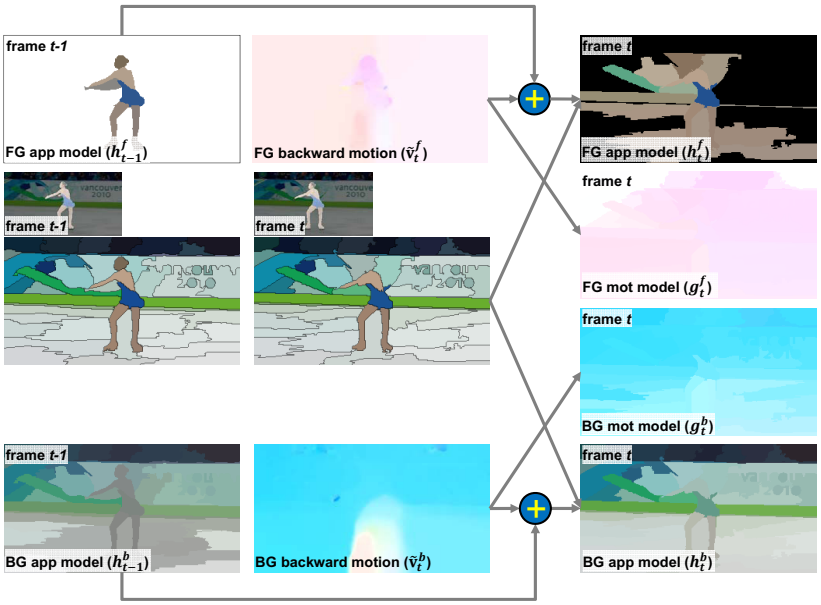

Fig. 3. Appearance model propagation by sequential Bayesian filtering. The models are constructed for individual superpixels. Note that we do not propagate motion models over time but compute them independently in each frame.

$h^\ell_t(\mathbf{c}; i)$, where $\mathbf{c}$ is a random variable for color and label $\ell \in \{f, b\}$ is a segment label. The motion histogram learned for the ith superpixel with label ℓ is also a normalized histogram denoted by $g^\ell_t(\mathbf{v}; i)$, where $\mathbf{v}$ is a random variable for motion. Since label information in the new frame is not available, we simply transfer labels from the previous frame using pixelwise motion $\mathbf{v}^\ell_t$ and set initial label to each pixel. Note that we need to maintain appearance and motion models for foreground and background separately in each superpixel.

4.2 Propagation of Appearance Models

Maintaining accurate appearance models of foreground and background is critical to obtain reliable likelihoods of both labels in each pixel. For the purpose, we propagate density function for appearance model by sequential Bayesian filtering, which is composed of prediction and update steps given by

$$p(x_t|z_{1:t-1}) = \int p(x_t|x_{t-1})\, p(x_{t-1}|z_{1:t-1})\, dx_{t-1} \tag{3}$$

$$p(x_t|z_{1:t}) \propto p(z_t|x_t)\, p(x_t|z_{1:t-1}), \tag{4}$$

where $p(x_t|z_{1:t-1})$ and $p(z_t|x_t)$ are prior and measurement density function, respectively. In our context, the histogram propagated from the previous frame

(a) FG/BG mask (b) $\Gamma_t(\mathbf{x})$ (c) $\Delta_t(\mathbf{x})$

Fig. 4. Motion consistency mask for likelihood estimation. Given the foreground (white) and background (gray) mask in (a) and the backward motions, the uncovered region (colored as black) can be found as (b). The black pixels in (c) are those with large color inconsistency between the current frame and the warped frame by the motion field. Refer to the text for more detail.

corresponds to prior distribution, and the appearance histogram observed in the current frame is measurement distribution. We now describe how the two distributions are constructed in our algorithm.

The prior distribution at frame t is estimated by a weighted sum of appearance models in frame $t-1$. The temporally propagated histogram $h^{\ell}_{t-1|t}(\mathbf{c}; i)$ corresponding to label ℓ in the ith superpixel is obtained as follows:

$$h^{\ell}_{t-1|t}(\mathbf{c}; i) = \frac{1}{n^{\ell}(\mathcal{S}_i)} \sum_{\mathbf{x} \in \mathcal{S}_i, L(\mathbf{x})=\ell} h^{\ell}_{t-1}(\mathbf{c}; s_{t-1}(\mathbf{x} + \mathbf{v}^{\ell}_t(\mathbf{x}))), \tag{5}$$

where $n^{\ell}(\mathcal{S}_i)$ is the number of pixels with label ℓ in the ith superpixel $\mathcal{S}_i$, $s_t(\mathbf{x})$ returns superpixel index of pixel located at $\mathbf{x}$ in frame t, and $\mathbf{v}^{\ell}_t(\mathbf{x})$ denotes backward motion vector for label ℓ observed at pixel $\mathbf{x}$. This procedure is illustrated in Figure 3.

For the measurement distribution denoted by $o^{\ell}_t(\mathbf{c}; i)$, we construct a normalized histogram for each label based on pixel colors in $\mathcal{S}_i$ at frame t. Note that each pixel whose label is not ℓ in the ith superpixel contributes to all bins in the histogram equally. Finally, the posterior of appearance in the current frame is obtained by the product of prior and measurement distributions as

$$h^{\ell}_t(\mathbf{c}; i) = o^{\ell}_t(\mathbf{c}; i) \cdot h^{\ell}_{t-1|t}(\mathbf{c}; i). \tag{6}$$

Note that the posterior is normalized to sum to one after we compute likelihoods for foreground and background. Unlike the appearance model estimated by sequential Bayesian filtering, the motion model $g^{\ell}_t(\mathbf{v}; i)$ is not propagated, but built from the motion field of the current frame.

4.3 Likelihood Computation and Label Estimation

Each superpixel has two sets of appearance and motion models; one is for foreground and the other is for background. Given the models at frame t, we compute foreground and background likelihoods of each pixel $\mathbf{x}$, denoted by $p^{\ell}(\mathbf{x})$ for label $\ell \in \{f, b\}$, which are given by a weighted geometric mean of appearance likelihood $h^{\ell}_t(\mathbf{c}(\mathbf{x}); s_t(\mathbf{x}))$ and motion likelihood $g^{\ell}_t(\mathbf{v}(\mathbf{x}); s_t(\mathbf{x}))$ as

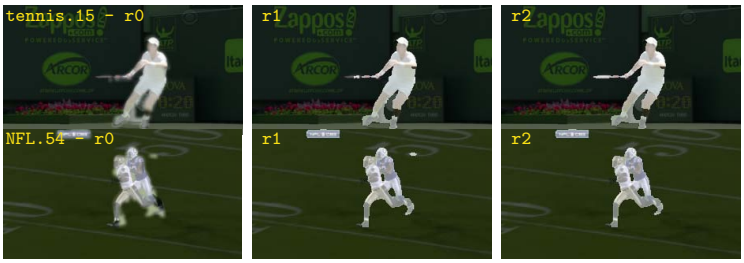

Fig. 5. Label update over iterations. **r0**: The initial label propagated from the previous frame. **r1**-**r2**: The updated label after the first and second iteration. The racket is recovered in *tennis* sequence, and the false foreground label on the ground in *NFL* sequence is removed.

$$p^{\ell}(\mathbf{x}) = h_t^{\ell}(\mathbf{c}(\mathbf{x}); s_t(\mathbf{x}))^{\alpha} \cdot g_t^{\ell}(\mathbf{v}(\mathbf{x}); s_t(\mathbf{x}))^{1-\alpha}, \tag{7}$$

where likelihoods are computed with the models in superpixel $s_t(\mathbf{x})$, and α ($0 \leq \alpha \leq 1$) controls relative weights between appearance and motion. If there is no foreground (or background) in a superpixel, the corresponding likelihood is set to zero.

If a pixel in the current frame is in the background region but its projected position by the backward background motion $\mathbf{v}_t^b(\mathbf{x})$ was in the foreground region in the previous frame, the pixel is likely to be uncovered from occlusion. The uncovered region can be identified by

$$\Gamma_t(\mathbf{x}) = L_{t-1}^f(\mathbf{x} + \tilde{\mathbf{v}}_t^b(\mathbf{x})) \wedge \tilde{L}_t^b(\mathbf{x}), \tag{8}$$

where $\tilde{L}_t^b(\mathbf{x})$ is the warped background mask at frame t and $L_{t-1}^f(\mathbf{x})$ is the foreground mask at frame $t-1$ as defined in Section 3. In motion likelihood estimation, the contribution of uncovered pixels for the foreground model is discarded.

Also, the pixel color consistency between the current frame and the warped image by the label's motion is checked, and pixels with large color inconsistency are ignored in the motion likelihood computation for the label, *i.e.*,

$$\Delta_t(\mathbf{x}) = \begin{cases} 1 \text{ if } \| \mathbf{c}_t(\mathbf{x}) - \mathbf{c}_{t-1}(\mathbf{x} + \mathbf{v}_t^{\ell}(\mathbf{x})) \|^2 > \theta \\ 0 \text{ otherwise} \end{cases}, \tag{9}$$

where $\mathbf{c}_t(\mathbf{x})$ and $\mathbf{c}_{t-1}(\mathbf{x})$ are the color values at $\mathbf{x}$ in the current and the previous frame. Figure 4 shows the examples of the $\Gamma_t(\mathbf{x})$ and $\Delta_t(\mathbf{x})$ masks, in which the black pixels are not used in motion likelihood computation.

Once foreground and background likelihoods of each pixel is computed, we estimate the label of each pixel by inference in Markov Random Field. Let $G = (V, E)$ be a graph with a set of vertices and edges, which are denoted by V and E, respectively. Each pixel corresponds to a vertex in the graph and edges connect four neighborhood vertices. Our objective is to minimize an energy function, which is composed of two terms—data and smoothness terms, which are given by observation potentials of individual vertices, $\Phi(\mathbf{x})$, and compatibility potentials of individual edges, $\Psi(\mathbf{x}, \mathbf{x}')$, respectively.

Fig. 6. Visualization of the foreground/background appearance models. Note that the proposed temporal propagation scheme maintains accurate appearance models, even where the background scene is occluded by foreground objects. Gray areas in the images for background appearance denote the absence of appearance models, and our algorithm learns the models quickly using new observations.

The observation potentials for foreground and background of each pixel $\mathbf{x}$ are given by

$$\Phi^{\ell}(\mathbf{x}) = \frac{p^{\ell}(\mathbf{x})}{p^{f}(\mathbf{x}) + p^{b}(\mathbf{x})}, \tag{10}$$

where $p^{f}(\mathbf{x})$ and $p^{b}(\mathbf{x})$ are computed by Eq. (7). The compatibility potential for an edge is defined by color difference between two adjacent pixels, $\mathbf{x}$ and $\mathbf{x}'$, which is given by

$$\Psi(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-||\mathbf{c}(\mathbf{x}) - \mathbf{c}(\mathbf{x}')||^2}{2\sigma_c^2}\right), \tag{11}$$

where σ_c is the parameter to control the effect of color difference. The optimization problem can be solved efficiently by binary belief propagation, and the labels are determined by comparing the believes for foreground and background at each pixel.

4.4 Iterative Update of Models and Labels

If the label of each pixel is re-estimated, the propagated appearance models and the estimated motion models in each superpixel should also be updated. Then, model propagation and label estimation procedures in Section 4.2 and 4.3 need to be repeated until convergence to improve overall performance. The foreground/background label estimation result in each iteration is presented in Figure 5, which shows gradual improvement of labels in each iteration.

Figure 6 illustrates the learned and propagated foreground and background models. Note that the initially occluded background regions (visualized as gray) are filled with correct appearance models using the information propagated from the previous frames.

Table 1. The experimental setup for comparison.

Algorithms	Block	Motion	Density estimation
Proposed method	ERS (300),	FG/BG layered motions,	histogram
– with grid blocks	grid (300),	FG/BG layered motions,	histogram
– with a single motion	ERS(300),	single motion,	histogram
Kwak *et al.* [13]	grid,	single motion,	KDE
Lim *et al.* [14]	grid,	single motion,	histogram

5 Experiment

We tested the proposed algorithm extensively with many videos involving various challenges such as background clutter, fast motion, occlusion, complex foreground shape, etc. Also, our algorithm is compared with the state-of-the-art techniques qualitatively and quantitatively.

5.1 Experiment Setup

Our algorithm requires two important external components: motion estimation and superpixel segmentation. Dense optical flow maps are estimated by the algorithm in [19], and superpixel segmentation is obtained from [18]. Although their results may affect overall performance of our algorithm substantially, we do not investigate their performance in this paper.

We evaluated the proposed algorithm called Generalized Background Subtraction using Superpixels (GBSSP), and two state-of-the-art algorithms developed by Kwak *et al.* [13] and Lim *et al.* [14]. The two algorithms [13, 14] employ block-based modeling and propagation strategy, where a regular rectangular blocks are used without sophisticated estimation of region boundaries. Table 1 summarizes the similarities and differences between the compared algorithms. Trajectory-based online moving camera background subtraction technique [11] has different characteristics compared to density propagation-based algorithms including ours; it assumes a certain camera model to classify the trajectories into foreground or background while ours does not have any assumption about parametric motion model. Consequently, it may show completely different performance depending on the choice of input sequences[2].

Our algorithm involves several free parameters[3]. Note that we fixed the parameters for each algorithm in all experiments to make the our evaluation fair and realistic. The initializations of all compared algorithms could be computed by motion segmentation followed by a few iterations of individual algorithms, which requires substantial amount of efforts for parameter tuning. Instead, to focus on the performance of foreground/background model propagation in all algorithms, we used the ground-truth labels for the first frames.

[2] Thus direct comparison with [11] is not conducted in our experiment given the situation that the implementation is not available. Please refer to the paper for indirect performance evaluation for a few common sequences.

[3] In Section 3, $r = 7$, and in Section 4.3, $\alpha = 0.7$, $\theta = 0.025$, $\sigma_c = 5$.

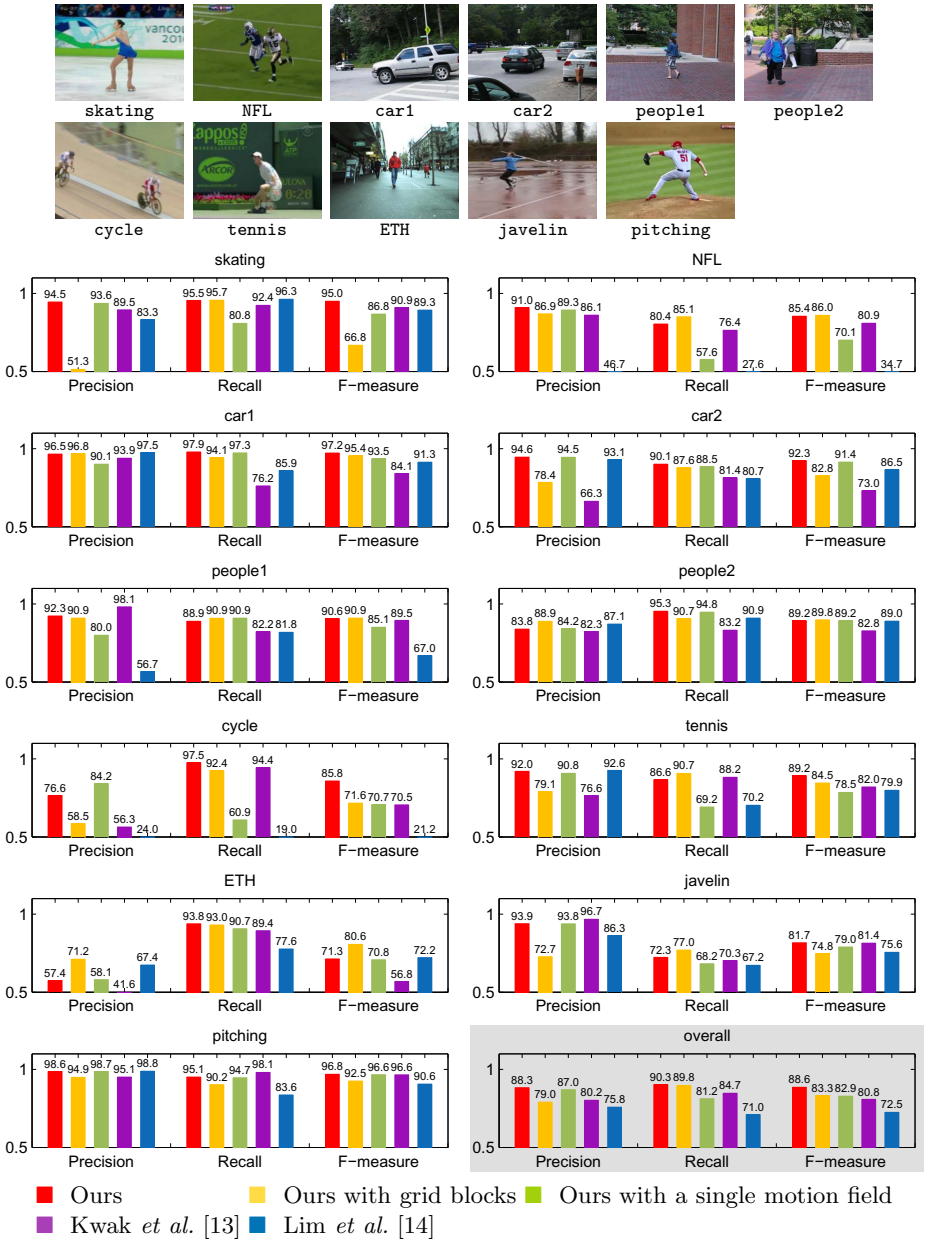

Fig. 7. Quantitative comparison results. The first frames of the tested sequences are illustrated at the top. The first 11 plots illustrate the precision and recall scores together with F-measure scores for the 5 different algorithms. Overall, the proposed algorithm outperforms [13] and [14] as illustrated in the highlighted graph. The benefit of superpixel segmentation and separate foreground/background motion are also supported by the results.

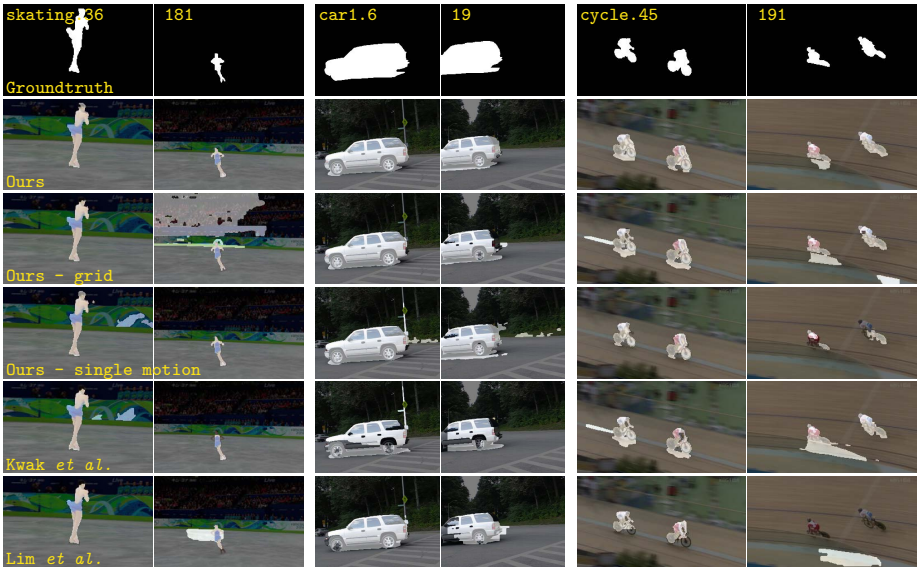

Fig. 8. Comparison with two internal and two external algorithms for *skating, car1*, and *cycle* sequences. The images in each row show the result of the five algorithms. See the text for discussion. **(Row1)** Groundtruth **(Row2)** Our algorithm **(Row3)** Ours with grid blocks **(Row4)** Ours with a single motion field **(Row5)** Kwak *et al.* [13] **(Row6)** Lim *et al.* [14]

For the performance evaluation of our algorithm, we selected 11 challenging videos, which include *car1, car2, people1* and *people2* from the Hopkins 155 dataset [20], *skating* and *cycle* from [13], *NFL* and *tennis* from [14], *ETH* from [21], *javelin* and *pitching*. Some sequences, *cycle*, *skating*, or *NFL*, contain several hundred frames and involve significant appearance changes in both foreground and background. The first frames of all the tested sequences are shown at the top of Figure 7.

5.2 Performance Evaluation

We present the qualitative and quantitative performance of our algorithm (GB-SSP) compared to Kwak *et al.* [13], and Lim *et al.* [14]. In addition, two variations of our algorithm are tested; one is our algorithm with a single motion field and the other is based on grid blocks instead of superpixels.

For quantitative comparison, precision and recall scores are computed based on the labels generated by the algorithms and manually annotated ground-truths. We used the precision and recall measures defined in [11]:

$$precision = \frac{TP}{TP + FP} \quad \text{and} \quad recall = \frac{TP}{TP + FN},$$

where TP, FP, TN, and FN denote the number of true-positive, false-positive, true-negative, and false-negative pixels, respectively. Note that the definition

Fig. 9. Comparison with two external algorithms for *NFL, people1, people2, pitching, car2, ETH, tennis.* and *javelin* sequences. **(Row1)** Groundtruth **(Row2)** Our algorithm **(Row2)** Kwak *et al.* [13] **(Row3)** Lim *et al.* [14]

of precision is different from the one used in [13, 14], $\frac{TN}{FP+TN}$, which tends to exaggerate precision performance when background area is large. The precision used here is more discriminative especially when the foreground region is small, and it is more consistent with human perception.

Figure 7 illustrates the quantitative results from the five algorithms for all the eleven test sequences. F-measure, which is the harmonic mean of precision and recall, for each algorithm is shown together with the precision and recall value. GBSSP is particularly better than [13] and [14] in *cycle* and *tennis* sequence, and achieves considerable performance improvement over the other two methods in many cases. Overall, our algorithm outperforms [13] and [14] by about 8% and 16%, respectively, on average. The accuracy of [13] is relatively low in *car1, car2* and *cycle* sequences and the performance of [14] is even worse; it fails in *cycle* and *people1*. It is notable that both the construction of superpixels and the estimation of separate motion fields are helpful to improve performance and that the combination of two components even boosts performance.

Figure 8 shows the foreground/background segmentation results in *skating*, *car1*, and *cycle* sequences for all five algorithms. Although they involves severe motion blur, low contrast background, or non-planar geometry, our method

performs very well on most sequences. It gets lower precision scores than human perception in some sequences since their ground-truth marking does not include the cast shadows by foreground objects, which are easily classified as foreground. By explicitly considering the uncovered region and the pixel consistency, which is made possible by the separate motion field estimation, the foreground mask does not bleed to nearby background areas with similar colors.

We provide more comparisons with the two state-of-the-art algorithms in Figure 9. Our algorithm illustrates visually better or at least similar results in all sequences compared to all other methods. *ETH* sequence is apparently most challenging, which is probably because foreground objects appear as tiny blobs and the facade color is very similar to the pedestrians. All algorithms fail to produce satisfactory results in this sequence.

Since our algorithm does not require complex inference procedures such as nonparametric belief propagation and sequential Bayesian filtering based on Gaussian mixture models, it is 6∼7 times faster than [13]. In a standard laptop computer, it approximately takes 6 seconds per frame.

6 Conclusion

We presented an online foreground/background segmentation algorithm for videos captured by a moving camera. Our algorithm maintains reliable foreground and background appearance models over time and obtains the label of each pixel based on the learned appearance and motion models. For the purpose, it performs superpixel segmentation in each frame and computes foreground/background motion fields by exploiting segmentation mask. The appearance models of each superpixel are propagated through a sequential Bayesian filtering and the motion models are also estimated for each superpixel. Pixelwise foreground and background likelihoods are computed by the appearance and motion models, and binary belief propagation is applied to infer the labels in each iteration. This procedure is performed multiple iterations in each frame, and the final labels are obtained upon convergence. Our algorithm is conceptually simple and presents significant performance gain compared to the state-of-the-art techniques.

Acknowledgements. We thank the reviewers for valuable comments and suggestions. The work is supported partly by the ICT R&D programs of MSIP/KEIT (No. 10047078), MKE/KEIT (No. 10040246), and MSIP/IITP [14-824-09-006, Novel computer vision and machine learning technology with the ability to predict and forecast; 14-824-09-014, Basic software research in human-level lifelong machine learning (Machine Learning Center)].

References

1. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
2. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR, pp. 246–252 (1999)
3. Lee, D.: Effective gaussian mixture learning for video background subtraction. IEEE TPAMI 27, 827–832 (2005)
4. Sheikh, Y., Shah, M.: Bayesian object detection in dynamic scenes. In: CVPR (2005)
5. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE TPAMI 19, 780–785 (1997)
6. Han, B., Davis, L.: Density-based multi-feature background subtraction with support vector machine. IEEE TPAMI 34, 1017–1023 (2012)
7. Hayman, E., Eklundh, J.O.: Statistical background subtraction for a mobile observer. In: ICCV (2003)
8. Mittal, A., Huttenlocher, D.: Scene modeling for wide area surveillance and image synthesis. In: CVPR (2000)
9. Yuan, C., Medioni, G., Kang, J., Cohen, I.: Detecting motion regions in the presence of a string parallax from a moving camera by multiview geometric constraints. IEEE TPAMI 20 (2007)
10. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV (2009)
11. Elqursh, A., Elgammal, A.: Online moving camera background subtraction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 228–241. Springer, Heidelberg (2012)
12. Sand, P., Teller, S.: Particle video: Long-range motion estimation using point trajectories. In: CVPR, pp. 2195–2202 (2006)
13. Kwak, S., Lim, T., Nam, W., Han, B., Han, J.H.: Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering. In: ICCV, pp. 2174–2181 (2011)
14. Lim, T., Han, B., Han, J.H.: Modeling and segmentation of floating foreground and background in videos. Pattern Recognition 45, 1696–1706 (2012)
15. Cui, X., Huang, J., Zhang, S., Metaxas, D.N.: Background subtraction using low rank and group sparsity constraints. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 612–625. Springer, Heidelberg (2012)
16. Ochs, P., Brox, T.: Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In: ICCV, pp. 1583–1590 (2011)
17. Liu, C., Freeman, W., Adelson, E., Weiss, Y.: Human-assistied motion annotation. In: CVPR (2008)
18. Liu, M.Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: CVPR, pp. 2097–2104 (2011)
19. Liu, C.: Beyond pixels: exploring new representations and applications for motion analysis. PhD thesis, Massachusetts Institute of Technology (2009)
20. Tron, R., Vidal, R.: A benchmark for the comparison of 3-d motion segmentation algorithms. In: CVPR (2007)
21. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: ICCV 2007 (2007)