Visual-Attention-Based Background Modeling for Detecting Infrequently Moving Objects

Yuewei Lin, *Student Member, IEEE*, Yan Tong, *Member, IEEE*, Yu Cao, *Member, IEEE*, Youjie Zhou, *Student Member, IEEE*, and Song Wang, *Senior Member, IEEE*

***Abstract*— Motion is one of the most important cues to separate foreground objects from the background in a video. Using a stationary camera, it is usually assumed that the background is static, while the foreground objects are moving most of the time. However, in practice, the foreground objects may show *infrequent motions*, such as abandoned objects and sleeping persons. Meanwhile, the background may contain frequent local motions, such as waving trees and/or grass. Such complexities may prevent the existing background subtraction algorithms from correctly identifying the foreground objects. In this paper, we propose a new approach that can detect the foreground objects with frequent and/or infrequent motions. Specifically, we use a visual-attention mechanism to infer a complete background from a subset of frames and then propagate it to the other frames for accurate background subtraction. Furthermore, we develop a feature-matching-based local motion stabilization algorithm to identify frequent local motions in the background for reducing false positives in the detected foreground. The proposed approach is fully unsupervised, without using any supervised learning for object detection and tracking. Extensive experiments on a large number of videos have demonstrated that the proposed approach outperforms the state-of-the-art motion detection and background subtraction methods in comparison.**

***Index Terms*— Infrequently moving objects, local motion stabilization, object detection, visual attention.**

I. Introduction

IN MANY video surveillance tasks, it is necessary to separate foreground objects of interest, which can be persons, vehicles, animals, and so forth, from the background [1]. Based on the extracted foreground objects, high-level tasks, such as detecting/tracking target objects [2] and recognizing activities from videos, can be addressed more effectively. Assuming that the camera is stationary, motion plays a key role in video-based foreground/background separation: foreground objects are usually moving, while the background is relatively static. Many approaches, such as optical flow and background subtraction, have been developed to detect the motions of foreground objects, based on which the foreground and the background can be separated.

Optical flow requires that the foreground objects move all the time. However, in practice, foreground objects may show *infrequent motions*, i.e., objects remain static for a long time and have (short duration) motions occasionally, e.g., abandoned objects [3], removed objects [4], and persons stopping for a while and then walking away [5], [6]. As an example shown in Fig. 1, a red duffle bag was moving with a person at the beginning of the video and then was abandoned on the grassland for the rest of the video. Detection of such an unattended bag is of particular importance in surveillance. However, as shown in the second row of Fig. 1, optical flow fails to detect the bag when it stays stationary.

Background subtraction is another type of effective approach that has been widely used to detect the moving foreground objects from a clean background [7]–[12]. Its basic idea is to estimate a clean background image (without any foreground objects) and then calculate, pixelwise, the difference between each video frame and the background image. Assuming that the appearance difference is significant between the foreground and the background, the regions with large appearance difference are detected as the foreground and the remaining regions are treated as the background [13]. However, it is also difficult to detect *infrequently moving objects* using the existing background subtraction approaches. The major difficulty is to estimate the background image: the infrequently moving objects stay stationary for most of the time, and thus could be easily taken as part of the background, as shown in the third row of Fig. 1. More seriously, the background may not be absolutely static in the video. Other than camera shake, the scene itself may contain frequent local motions, such as trees/grass waving in the breeze, which could be easily confused as the foreground.

In this paper, we propose a fully unsupervised approach to identify foreground objects with frequent and/or infrequent motions. In this paper, we consider the cases in which the camera is mostly stationary, while having few abrupt movements. In this approach, we first divide a long streaming video into subvideos (called super-clips later in this paper) so that the background in each subvideo does not show significant change. Within each super-clip, we develop algorithms to identify *regions of difference* (RoD) between temporally nearby

Manuscript received July 5, 2015; revised October 21, 2015; accepted January 15, 2016. Date of publication February 8, 2016; date of current version June 5, 2017. This work was supported in part by the National Science Foundation under Grant IIS-1017199, in part by the Army Research Laboratory under Grant W911NF-10-2-0060 (DARPA Mind's Eye Project), and in part by the National Science Foundation under CAREER Award IIS-1149787. This paper was recommended by Associate Editor K.-K. Ma.

Y. Lin, Y. Tong, Y. Zhou, and S. Wang are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208 USA (e-mail: ywlin.cq@gmail.com; tongy@cec.sc.edu; zhou42@email.sc.edu; songwang@cec.sc.edu).

Y. Cao is with the IBM Almaden Research Center, San Jose, CA 95120 USA (e-mail: caoy@us.ibm.com). He participated in this work while he was earning the Ph.D. degree at the University of South Carolina, Columbia, SC, USA. His contact information refers only to his affiliation rather than any intellectual property.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2527258

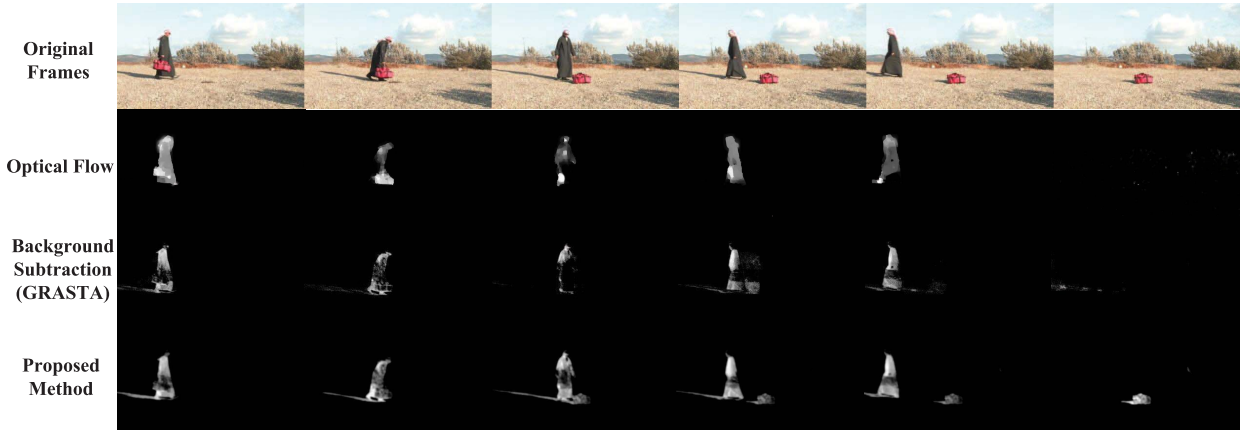


Fig. 1. Illustration that shows how the proposed method can catch the red duffle bag with infrequent motions. For comparison, we also include the results from a high-accuracy version of optical flow [14] and GRASTA [15], an online-discriminative-learning-based background subtraction method. Both optical flow and GRASTA can detect only the moving objects and fail to detect the red duffle bag after it was left on the grassland. Best viewed in color.

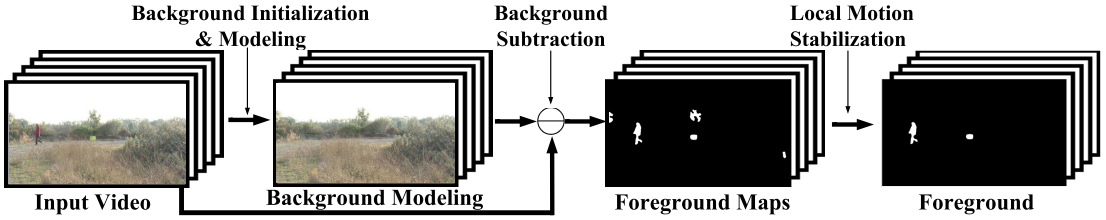


Fig. 2. Flowchart of the proposed method for foreground detection. Best viewed in color.

frames. Since there is no object-specific information of the foreground objects, a visual-attention mechanism is employed for identifying an RoD to be part of either an object or the background by assuming that the foreground objects should be more salient than the background. The RoDs that are identified as background regions are then propagated back-and-forth in the super-clip to construct complete background images, i.e., background models. With a complete background image for each frame, we can conduct background subtraction to identify the moving foreground objects. To address the local frequent motions in the background, we further develop a feature-matching-based local motion stabilization algorithm that can reduce the foreground false positives in background subtraction.

There are three major contributions in this paper.

1) A visual-attention-analysis-based algorithm is developed to evaluate whether an RoD shows the background in a frame.
2) A forward/backward background propagation algorithm is developed to construct complete background images.
3) A feature-matching-based local motion stabilization algorithm is proposed to suppress frequent local motions in the background and reduce false positives in foreground detection.

Our overall framework of foreground detection is illustrated in Fig. 2. The proposed method has been evaluated extensively on a large amount of data that contain objects with infrequent motions: 18 long videos (580 041 frames in total) from defense advanced research projects agency (DARPA) Mind's Eye project Y2 data set containing significant illumination changes, cluttered background, and motions in the scene, and six videos (18 650 frames in total) from the category of intermittent object motion in the ChangeDetection data set [5], [6]. The experiment results have demonstrated that the proposed method outperforms several state-of-the-art motion detection methods, especially with the infrequent moving foreground objects.

The rest of this paper is organized as follows. The related work is briefly discussed in Section II. In Section III, we introduce the proposed background-modeling method for constructing complete background images. In Section IV, we introduce the feature-matching-based local motion stabilization method for effective background subtraction. Section V reports the experimental results, followed by a conclusion in Section VI.

II. Related Work

Background subtraction may be the simplest approach for foreground detection [7], [9], [16]. The basic idea is to obtain a background image that does not contain any object of interest. Then, a video frame will be compared with the background image for foreground object detection [13]. The most critical and challenging task in background subtraction is background modeling, i.e., obtaining a clean background image, which generally includes background initialization and updating. Here, we give a brief review of this topic. See [17]–[19] for a comprehensive survey.

Assuming that the foreground objects have a color or intensity distribution different from that of the background, the majority of background-modeling approaches learn a background distribution at each pixel location, which is then used to classify each pixel in a video frame as background or foreground. The background distribution at each pixel

can be modeled parametrically, such as a Gaussian mixture model [7], or nonparametrically [20], such as kernel density estimation [21]. More recently, statistical background modeling has been extended to estimate the background distribution in a spatial or spatiotemporal neighborhood [22]–[25]. Sheikh and Shah [22] challenged the idea of modeling background distribution at each pixel and employed the correlation between spatially proximal pixels. Narayana *et al.* [23] proposed a kernel estimate at each pixel using data samples extracted from its spatial neighborhood in previous frames. Moshe *et al.* [24] directly modeled the statistics from 3D spatiotemporal video patches to capture both the static and the dynamic information of the scene. Hofmann *et al.* [26] proposed a pixel-based adaptive segmentator, which used a history of N background values to construct the background model and a random update rule. Hernandez-Lopez and Rivera [25] proposed to regularize the likelihood of each pixel belonging to background or foreground based on a quadratic Markov measure field model. However, this method [25] assumes that the first frame of the video does not contain the foreground and thus, cannot handle the case that the foreground objects are present at the beginning of the video. Shimada *et al.* [27] proposed a bidirectional background-modeling approach based on case-based reasoning whereby a background model was retrieved from an online constructed background database. Wang *et al.* [28] proposed to fuse the motion detection based on spatiotemporal tensor formulation and the foreground and background-modeling scheme based on split Gaussian models. Wang and Dudek [29] modeled background for each pixel by using a number of background values, followed by a classification process based on matching the background model templates with the current pixel values.

Besides modeling the statistics, foreground/background separation can be performed through low-rank subspace separation. Cui *et al.* [30] proposed a model using both low-rank and group sparsity constraints, which represented two observations, i.e., background motion caused by orthographic cameras lies in a low-rank subspace and pixels belonging to one trajectory tend to group together, respectively. He *et al.* [15] introduced an online background-modeling algorithm, named Grassmannian Robust Adaptive Subspace Tracking Algorithm (GRASTA), for low-rank subspace separation of background and foreground from randomly subsampled data. Lin *et al.* [31] proposed to pursue low-rank subspace in the spatiotemporal domain. However, the low-rank constraint tends to treat the objects with infrequent motions as the background.

In addition to color or intensity information, local texture information has also been employed in background modeling. Liao *et al.* [32] employed the local texture information by using a scale-invariant local ternary pattern (SILTP), which is modified from local binary patterns (LBPs). Han *et al.* [33] integrated the histogram of SILTP features and color information in a blockwise background model. Liu *et al.* [34] extended the SILTP to spherical center-symmetric SILTP by integrating spatiotemporal statistics for background modeling with a pan–tilt–zoom camera. Kim *et al.* [35] proposed to use scale-invariant feature transform (SIFT) features to generate adaptive multihomography matrices, which are then used to compensate for the global camera motion to detect the moving objects under the moving camera. Yao and Odobez [36] combined the local textures represented by LBPs and color features.

However, there is a common assumption in the existing background-modeling algorithms that the background is more frequently visible than the foreground. As a result, they are more likely to treat an object with infrequent motions as part of the background. In this paper, we employ a visual-attention-analysis-based mechanism to explicitly deal with the foreground objects with infrequent motions.

There is another set of literatures focusing on detecting abandoned/removed objects. For example, Lin *et al.* [37], [38] propose two background models to handle abandoned objects. The long-term background model is updated slowly by using a large learning rate, while the short-term background model is updated fast. Thus, the abandoned objects can be detected through comparing background subtraction results using the long-term and short-term background models. However, the long-term background model will cause the ghosting artifacts. Since the long-term background model is still updating, the abandoned objects will be treated as background finally. Tian *et al.* [4] model the background by using three Gaussian mixtures to represent the background and changes in different temporal scales, which also suffers from ghosting artifacts. Maddalena and Petrosino [39] explicitly detect the stopped objects from the moving ones by counting the consecutive occurrences (i.e., detected as a foreground) of an object from a sequence of frames. However, this model cannot detect the removed objects since it employs the first frame to initialize the background.

The saliency detection has recently raised a great amount of research interest and has been shown to be beneficial in many applications [40]–[42]. Although saliency detection has been employed in foreground/background separation in a few early attempts [43], [44], we would like to emphasize that the proposed approach is totally different from these approaches. In [43], regions with high visual saliency are identified on each frame as foreground without considering any motion cue. In [44], spatiotemporal segments with high visual saliency are identified from a video as foreground. While this method [44] considers motion cue in evaluating the visual saliency, it will fail to detect an infrequently moving object once it stays static and generates no motions. In this paper, we do not directly use visual saliency to separate foreground and background. Instead, we identify RoDs and compare the saliency values of an RoD in different frames to help construct the complete background images. Together with a step of background propagation, our method can better detect infrequently moving objects. In addition, directly using saliency in each frame to distinguish the foreground and the background may work poorly when the background is highly textured—highly textured regions are usually considered to be salient in most visual-attention models [45]. The proposed method compares the relative saliency of a region across frames to distinguish the foreground and the background and can better identify the

highly textured background, as shown later in the experiments (see Fig. 7).

III. Background Modeling With Visual-Attention Analysis

Background modeling intends to construct a complete background image that does not contain any object of interest so that foreground objects can be detected through background subtraction. The proposed background-modeling method involves the estimate of RoDs between temporally nearby frames. In the following, we first introduce the operation of RoD estimation and then introduce the proposed background-modeling method.

A. RoD Estimation

In this paper, we take the following steps to estimate the RoDs between two frames. First, we calculate the absolute pixelwise difference between these two frames for the three channels in hue, saturation, and value (HSV) space, respectively, from which an overall difference-image denoted by DI can be computed as follows:

$$\mathrm{DI}(x, y) = \max(\mathrm{DI}_h(x, y), \mathrm{DI}_s(x, y), \mathrm{DI}_v(x, y)) \quad (1)$$

where $\mathrm{DI}(x, y)$ denotes the overall difference value at a pixel (x, y) in DI; $\mathrm{DI}_h(x, y)$, $\mathrm{DI}_s(x, y)$, and $\mathrm{DI}_v(x, y)$ are the absolute difference values at pixel (x, y) for H, S, and V channels, respectively. Note that we generate DI by pixelwise taking the maximal value from HSV color channels, which is inspired by the winner-take-all mechanism [46] in human vision system. Then, DI is binarized to an RoD-map denoted by RM as follows:

$$\mathrm{RM}(x, y) = \begin{cases} 1, & \mathrm{DI}(x, y) \geq \max(\eta_1, \eta_2 \times \max(\mathrm{DI})) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where max(DI) is the maximum value in DI. η_1 and η_2 are two control parameters—η_1 sets an absolute threshold, whereas η_2 sets an adaptive threshold relative to the difference image DI.[1] Finally, each connected region in RM is taken as an RoD. For convenience, we use $\mathrm{RoD}(f, g)$ and $|\mathrm{RoD}(f, g)|$ to represent the RoDs and the total area of the RoDs between two frames f and g.

B. Background Initialization

From this section, we introduce the proposed method for background modeling and background subtraction, starting from an input long streaming video. For a long streaming video, there may be intermittent, abrupt background changes, such as those caused by sudden illumination change or camera shake. In this paper, we first divide the long video into a set of *super-clips* so that each super-clip does not contain abrupt background change. In this way, we can perform background modeling for each super-clip independently.

Specifically, the super-clips are constructed as follows. First, the input long video is uniformly divided into short video clips $C_i, i \in \{1, 2, \ldots, M\}$, with a predefined length of N frames for each video clip. A *key frame* c_i is then selected from each video clip C_i as its representative. In this paper, we simply pick the middle frame $f_{i\lfloor N/2 \rfloor}$ as the key frame c_i for the ith clip $C_i = \{f_{ij}\}$, $j \in \{1, 2, \ldots, N\}$. Starting from clip C_1, the key frame c_1 is compared with each $c_i (i > 1)$ sequentially until reaching a clip C_p with $|RoD(c_1, c_p)|$ larger than a threshold, which we empirically choose to be half of the image area. We then merge all the clips $C_i (1 \leq i < p)$ into the first super-clip. The second super-clip is generated similarly starting from C_p. This process is repeated until it gets to the last clip C_M. The number of super-clips is further reduced by merging nonadjacent super-clips if their temporally nearest key frames are sufficiently similar, which is set to be true if the total area of their estimated RoDs is smaller than 20% of the image area. This merging process is very useful for the temporary background change, e.g., for outdoor videos, the illumination may get darker for a while and then get back to normal, and the super-clips before and after the illumination change can be merged into a longer super-clip. As shown in Fig. 4, the key frames with the same-colored bounding boxes belong to the same super-clip.

Each constructed super-clip consists of a sequence of nonoverlapped and fixed-length short video clips, each of which needs a background image to accommodate the possible slow background variations within the super-clip. For each video clip C_i, the key frame c_i, which is the middle frame of C_i in this paper, is employed as its initial background image b_i such that $b_i = c_i$. In the following section, we introduce a propagation algorithm to update the initial background image b_i for each C_i by identifying foreground regions from b_i and replacing them with underlying background regions found from other key frames.

C. Background Propagation Based on Visual-Attention Analysis

Within a super-clip, we assume that for each pixel, there is at least one key frame on which this pixel is located in the background. Our goal is to identify such background pixels from different key frames and then combine them to form a complete background image. In this paper, we identify RoDs between adjacent background images and use these RoDs, instead of individual pixels, for constructing complete background images. Let us consider a super-clip with m clips $C_i, i \in \{1, 2, \ldots, m\}$ and our approach consists of: 1) a forward propagation from C_1 to C_2, then from C_2 to C_3, until it gets to C_m, followed by 2) a backward propagation from C_m to C_{m-1}, then from C_{m-1} to C_{m-2}, until it gets back to C_1. For each clip C_i, we construct a background image b_i, which is initialized as the key frame c_i. Without loss of generality, let us consider one step of forward propagation, say from C_{i-1} to C_i, which only updates the background image b_i, as follows. Note that when performing this step of propagation, b_{i-1} is not the original key frame c_i. Instead, it has been updated with the finished propagations from C_1 up to C_{i-1}.

1) Calculating the RoDs between b_{i-1} and b_i.
2) For each RoD (connected region) R, let $b_{i-j}(R)$, $j \in \{1, 2, \ldots, k\}$, and $b_i(R)$ be the appearance of the region R on the updated background images b_{i-j}, and to be updated background image b_i, respectively.

[1] $\eta_1 = 0.1$ and $\eta_2 = 0.2$ are chosen empirically in our experiments.

3) Constructing k new candidate background images $b^R_{(i-j)\to i}, j \in \{1, 2, \ldots, k\}$ which are obtained by replacing the region R in b_i by using $b_{i-j}(R)$, respectively

$$b^R_{(i-j)\to i}(x, y) = \begin{cases} b_{i-j}(x, y), & \text{if } (x, y) \in R \\ b_i(x, y), & \text{otherwise.} \end{cases} \quad (3)$$

4) Calculating the *background likelihood* of region R on b_i and $b^R_{(i-j)\to i}$ and denote them as $P_i(R)$ and $P_{(i-j)\to i}(R), j \in \{1, 2, \ldots, k\}$, respectively. Obtaining the j^* which has the maximal $P_{(i-j)\to i}(R)$ by

$$j^* = \arg\max_{j\in\{1,\ldots,k\}} P_{(i-j)\to i}(R). \quad (4)$$

5) If $P_{(i-j^*)\to i}(R) > P_i(R)$, we update the background image b_i by using $b^R_{(i-j^*)\to i}$.
6) Otherwise, no update to b_i in terms of the region R (other RoDs between b_{i-1} and b_i may still update b_i).

$P_i(R)$ describes the likelihood that the region R is located in background in b_i. Similarly, $P_{(i-j)\to i}(R)$ describes the likelihood that the region R is located in background in $b^R_{(i-j)\to i}$. In this paper, we employ a visual-attention mechanism to examine whether a region catches people's attention as a foreground object usually does. Based on this, we define the background likelihood as

$$P_i(R) \propto \frac{1}{\mathrm{SV}_i(R)} \quad (5)$$

where $\mathrm{SV}_i(R)$ is the saliency of region R in the currently estimated background image b_i. To stimulate this mechanism, the saliency value of a region, which closely relates to human attention, is computed as the difference between this region and its spatial surrounding (also known as center-surround difference [47]). For example, considering a specific RoD R enclosed in a red contour, as illustrated in Fig. 3(b), we find its surrounding region S as the region between the red contour and a blue box. In this paper, we construct the blue box by doubling the height and the width of the rectangular bounding box around the red contour (R). In the following, we elaborate on the region saliency and this background propagation process.

1) Region Saliency: To compute the center-surround difference, we calculate the image statistics in R and S, respectively. Specifically, we derive the histogram in HSV color space and then employ the χ^2 distance over all the color channels to measure the center-surround difference

$$d_c(R, S) = \sum_q \frac{1}{2} \sum_{b\in 1,\ldots,N_{\mathrm{bin}}} \frac{\left(H^q_{R,b} - H^q_{S,b}\right)^2}{H^q_{R,b} + H^q_{S,b}}, \quad q \in \{h, s, v\} \quad (6)$$

where $H^q_{R,b}$ and $H^q_{S,b}$ denote the bth bin of the histograms of the q channel in the HSV color space for the regions R and S, respectively. N_{bin} is the number of bins in the histogram, which we empirically set as 32.

As a global measurement, the histogram-based distance ignores the spatial information of pixels in each region. It is suggested by [45] and [48] that pixels near the boundary between R and S are more important than the others for computing saliency. Thus, we also measure the local contrast between R and S along the contour of R as follows:

$$d_l(R, S) = \sum_l |\bar{x}_{lR} - \bar{x}_{lS}|, l \in \text{pixels along the contour of } R \quad (7)$$

where

$$\bar{x}_{lR} = \frac{\sum_{p\in N_l\cap R}(h_{lp} + s_{lp} + v_{lp})}{|N_l \cap R|} \quad (8)$$

$$\bar{x}_{lS} = \frac{\sum_{p\in N_l\cap S}(h_{lp} + s_{lp} + v_{lp})}{|N_l \cap S|} \quad (9)$$

where N_l denotes the neighboring region centered at the pixel l, which we set as a 7×7 window empirically; h_{lp}, s_{lp}, and v_{lp} represent the values in the HSV space for the pth pixel in N_l.

In this paper, we define the saliency of the region R in b_i by combining these two center-surround distances, as

$$\mathrm{SV}_i(R) = d_c(R, S) \times d_l(R, S). \quad (10)$$

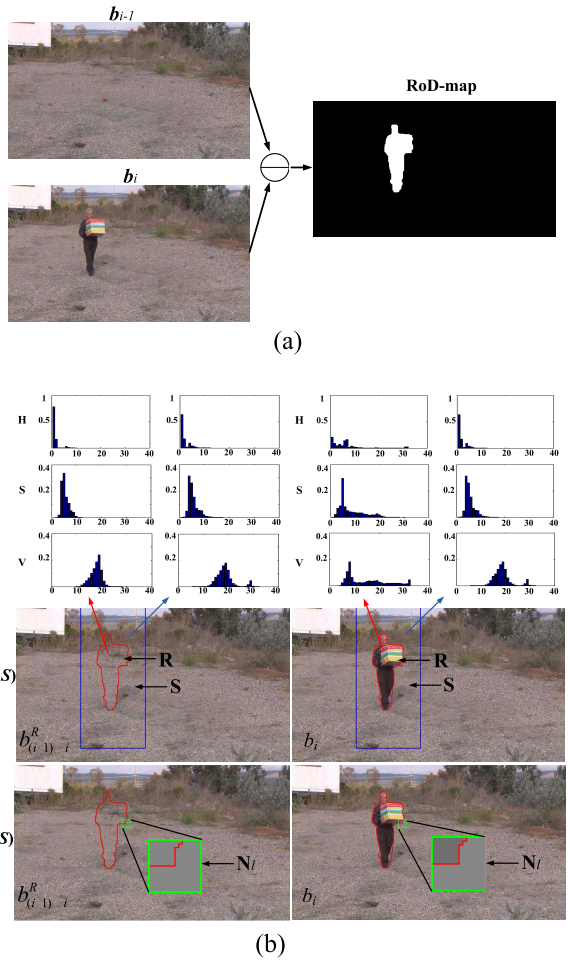


Fig. 3. Illustration of calculating region saliency. (a) RoD-map between the two initially estimated background images b_{i-1} and b_i. (b) Region saliency of R on $b^R_{(i-1)\to i}$ (left) and b_i (right) by combining two center-surround differences, with the histograms of R and S shown on the top. $d_c(R, S)$ takes the value of 0.0529 and 0.3356 on $b^R_{(i-1)\to i}$ and b_i, respectively, and $d_l(R, S)$ takes the value of 0.0316 and 0.0470 on $b^R_{(i-1)\to i}$ and b_i, respectively. Best viewed in color.

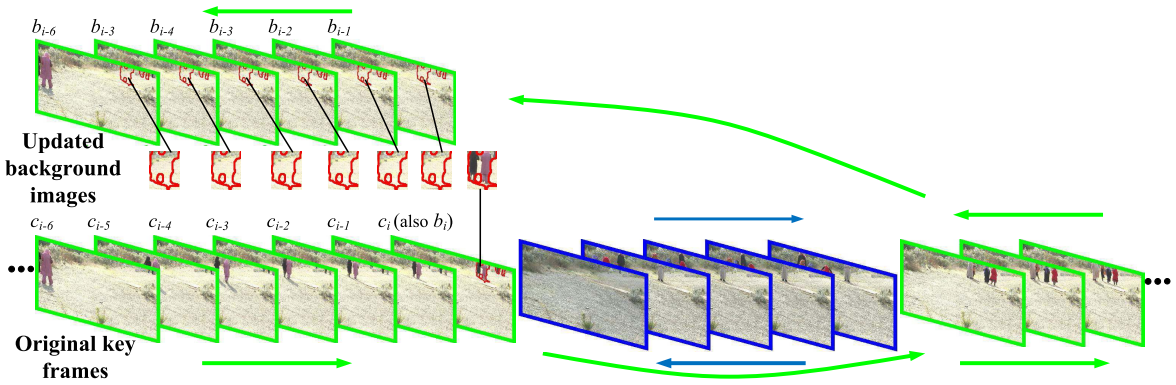


Fig. 4. Illustration of the forward and backward background propagation. Each image in the second row represents a key frame of a video clip, which is also considered as the initialized background image for the related clip. The key frames with the same color of bounding boxes belong to the same super-clip. Red contours: RoD R considered. Best viewed in color.

An example is shown in Fig. 3(b), where the region R shows much lower center-surround differences $d_c(R, S)$ and $d_l(R, S)$ on the candidate background image $b^R_{(i-1)\to i}$ [the left column of Fig. 3(b)] than on the initial estimated background image b_i [the right column of Fig. 3(b)]. In this case, we need to update the background image b_i in terms of the region R.

2) Background Propagation: By performing the forward propagation from C_1 up to C_m in the super-clip, we expect that the backgrounds shown on earlier clips will be propagated to the later clips. After that, we need to perform backward propagation from C_m down to C_1 since the foreground on earlier clips, such as C_1, cannot be replaced by backgrounds in the forward propagation. Let us still use the forward propagation from b_{i-1} to b_i as an example. As given in (3), we construct k candidate background images $b^R_{(i-j)\to i}$, $j \in \{1, 2, \ldots, k\}$. In this paper, we set k to be 6, i.e., we construct $b^R_{(i-j)\to i}, j \in \{1, 2, \ldots, 6\}$ by copying region R from updated background images $b_{i-1}, b_{i-2}, \ldots, b_{i-6}$. In this way, we calculate the region saliency of R on these six candidate background images and the background image b_i, then pick the one on which region R shows the lowest saliency (i.e., the highest background likelihood) to update b_i.

An example is shown in Fig. 4, an RoD R is shown on six updated background images and the to-be-updated background image b_i. The saliency value of R on the candidate background images $b^R_{(i-1)\to i}, b^R_{(i-2)\to i}, \ldots, b^R_{(i-6)\to i}$ are 0.0076, 0.0074, 0.0072, 0.0073, 0.0067, and 0.0063, respectively; whereas the saliency of R on b_i is 0.0639. As a result, in this step of propagation, we use the candidate background image $b^R_{(i-6)\to i}$ as the updated b_i, which can also be considered as replacing the region R in b_i by using the R in b_{i-6}. From Fig. 4, it can be seen that the persons appearing at the beginning of the super-clip (e.g., c_{i-6}) cannot be removed in forward propagation. To construct clean and complete background images, we have to perform backward propagation (from C_m down to C_1). These persons will be replaced by the background if they leave the original location at some later key frames. Note that the backward propagation is performed on the background images that have been updated in the forward propagation.

IV. Background Subtraction and Local Motion Stabilization

A. Background Subtraction

Once the background image is constructed for each video clip C_i, background subtraction can be conducted by subtracting every frame in the video clip from the background image. We use the same algorithm for calculating the RoDs (see Section III-A) for background subtraction. The only difference is that we input a frame and the background image, instead of two frames, to calculate the RoDs, which are taken as the detected foreground objects.

B. Local Motion Stabilization Based on Feature Matching

The pixelwise background subtraction as presented above is sensitive to frequent local motions in the scene (background), such as trees and/or grass waving in the breeze. As a result, the waving trees and/or grass will be misdetected as foreground objects. To suppress the effect of local motions in background subtraction, we propose a local motion stabilization method based on feature matching.

In this case, the detected RoDs from background subtraction (i.e., subtracting a frame f to the background image b) may come from the foreground objects or the background local motions. We examine each RoD R in f and identify it to be part of the foreground or the background. Our basic idea is that, if R is part of the background in f, then $f(R)$, the region R in f, and $b(R)$, the region R in b, should share a lot of appearance features, such as SIFT features, although there is background local motion between f and b. The SIFT features are invariant to image scale and rotation, and robust to changes in illumination, noise, and minor changes in viewpoint. Since SIFT features are invariant to image scale and rotation, robust to changes in illumination and noise, and have the highest matching accuracy compared with other local features [49], we detect and match the SIFT features between $f(R)$ and $b(R)$ and define the background likelihood of R in f as

$$\Omega(f(R)) = \frac{N_{\text{matched}}}{\max(N_f, N_b)} \tag{11}$$

TABLE I
INFORMATION OF IMAGE DATA SETS USED IN EXPERIMENTS

Datasets	Number of Videos	Frame Size	Number of Frames			Ground Truth Labels
			Min.	Max.	Total	
DARPA Dataset	18	1280×720	19,435	55,776	580,041	588,902 bounding boxes
DARPA Clips	20	1280×720	1,000	1,000	20,000	40,594 bounding boxes
ChangeDetection	6	432×288/320×240	2,500	4,500	18,650	pixelwise labels

Fig. 5. Illustration of the proposed local motion stabilization algorithm based on SIFT matching. Left: current frame. Right: background image. The true foreground object (a person) is detected because of a few pairs of matched feature points. The false alarm (a waving tree) can be removed due to a lot of matched feature points. Best viewed in color.

where N_{matched} denotes the number of matched SIFT pairs [50][2] between $f(R)$ and $b(R)$; N_f and N_b denote the number of detected SIFT feature points on $f(R)$ and $b(R)$, respectively. In our experiments, if $\Omega(f(R))$ is larger than a predefined threshold τ,[3] R is considered to be part of the background in f, and we remove it from the foreground detection result.

An example of local motion stabilization using the SIFT matching is illustrated in Fig. 5. The left image represents a frame f, and the right image represents the background image b constructed, as described in Section III. After background subtraction, two RoDs, shown as the regions enclosed in the red contours, are detected. One of them contains a real object (i.e., a person); and the other one is part of background (i.e., a waving tree). The SIFT matching between the real object and the background returns a few pairs of matched points (top-left subimages); while it returns a lots of matched pairs between the trees in these two images (top-right subimages).

Although there is a risk of missing foreground objects, if the foreground objects have very similar appearance as the background, it is worth applying the motion stabilization to reduce a large number of false positives due to frequent local motions present in the background, especially in an outdoor environment. In practice, the proposed method can work well even if the foreground objects have similar textures as the background. As demonstrated in Fig. 7, waving tree/grass can be effectively eliminated from foreground detections; while the soldiers in the camouflage uniforms, which have similar texture as the bushes, are kept using the proposed stabilization algorithm.

[2]In our work, we use the code from http://www.cs.ubc.ca/~lowe/keypoints/

[3]τ is set to 0.1 empirically in our experiments.

V. EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness of our method, we have conducted extensive experiments on two data sets: DARPA data set and ChangeDetection data set [5], [6]. The detailed pieces of information of these data sets are listed in Table I.

The performance of the proposed method is compared with five state-of-the-art background subtraction methods including the method based on LBP and color features [36], a method based on mean-shift [51], the visual background extractor (ViBe) method [13], the GRASTA method [15], and the spatial coherence self-organizing background subtraction (SC-SOBS) [52]. For these methods used for comparison, we use the codes provided by their authors.

The proposed method and the other methods in comparison are evaluated quantitatively in terms of *Recall*, *Precision*, and *F1-measure*. In this paper, we define *Recall* as the ratio of the overlapped area between the ground truth bounding boxes (or foreground regions) and the detected foreground regions to the area of the ground truth bounding boxes (or foreground regions)[4]; and define *Precision* as the ratio of the overlapped area between the ground truth bounding boxes (or foreground regions) and the detected foreground regions to the area of the detected foreground regions. The *F1-measure* is defined as the harmonic mean of *Precision* and *Recall*, i.e., $F1 = 2 \times (\text{Precision} \times \text{Recall}/\text{Precision} + \text{Recall})$.

A. Experimental Results on the DARPA Data Set

1) Long Streaming Videos: In the first set of experiments, we evaluate the proposed method on the DARPA data set, which has a total number of 580 041 frames and is a subset of the DARPA Mind's Eye project Y2 data set. Specifically, 18 videos with manually annotated ground truth are selected. Each video is taken from a fixed camera viewpoint and contains significant local motions (trees and/or grass waving) and illumination changes in the scene. Ground truth is bounding-box based, with 588 902 bounding boxes in total.[5] The validation studies on this data set intend to demonstrate that the proposed method is capable of handling challenging scenes and can be scaled up to deal with large data.

We first evaluate the overall performance of foreground detection on the entire DARPA data set. To justify the use of the two center-surround distances described in Section III-C1, we also report in Table II the performance of the proposed method without considering either the color-histogram-based distance $d_c(R, S)$ or the local-contrast-based distance $d_l(R, S)$

[4]Because the ground truth labels in DARPA are given by bounding boxes, the score of *Recall* tends to be low when evaluated on these two data sets.

[5]In this paper, we have manually corrected some incorrect labels.

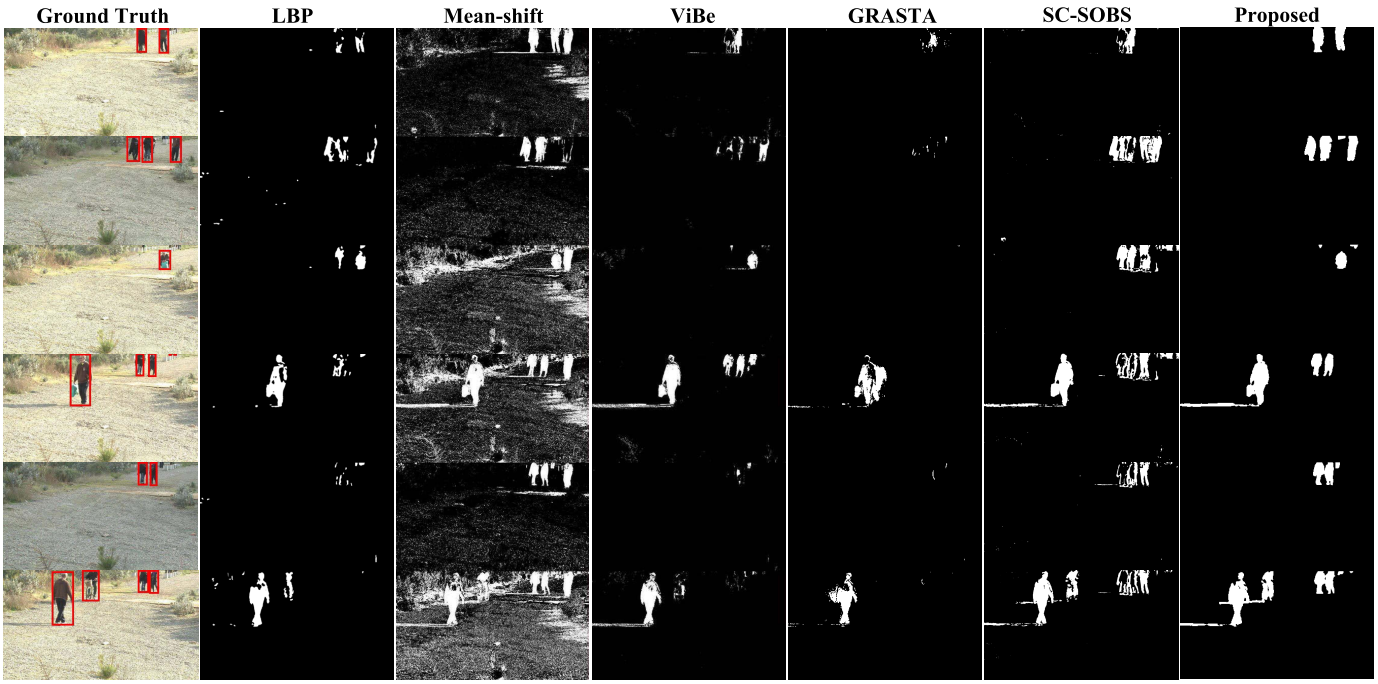


Fig. 6. Detection results on one sample video in the DARPA data set. The first column shows the original images from the video, and the following columns from left to right show the background subtraction results using the LBP-based method [36], the mean-shift-based method [51], the ViBe method [13], the GRASTA method [15], and the proposed method, respectively. Top to bottom: frames 60, 2200, 11000, 13800, 16350, and 21200. Best viewed in color.

TABLE II

PERFORMANCE COMPARISON ON THE DARPA DATA SET. VARIANTS OF THE PROPOSED METHOD WITH DIFFERENT COMPONENTS: w/o C-HIST AND w/o LC DENOTE THE PROPOSED METHOD WITHOUT THE COLOR-HISTOGRAM-BASED DISTANCE AND WITHOUT THE LOCAL-CONTRAST-BASED DISTANCE, RESPECTIVELY, AND w/o STABILIZATION DENOTES THE PROPOSED METHOD WITHOUT THE LOCAL MOTION STABILIZATION

Model	Recall	Precision	F1-measure
LBP based [36]	0.4756	0.5211	0.4973
Mean-shift based [51]	0.5673	0.0649	0.1165
ViBe [13]	0.4445	0.5425	0.4886
GRASTA [15]	0.2738	0.6545	0.3861
SC-SOBS [52]	0.5238	0.5935	0.5565
Proposed (w/o C-hist)	0.5259	0.6437	0.5789
Proposed (w/o LC)	0.5601	0.6662	0.6086
Proposed (w/o stabilization)	**0.5718**	0.6555	0.6108
Proposed	0.5701	**0.6877**	**0.6234**

in measuring the saliency. We can see that both of these two distances contribute to the performance of the proposed method, although the color-histogram-based distance contributes more to the final performance than the local-contrast-based distance does.

To demonstrate the effectiveness of the SIFT-matching-based local motion stabilization, the performances of the proposed method with and without local motion stabilization are compared. As shown in Table II, the proposed method with local stabilization significantly outperforms all the other methods in comparison in terms of *Precision* and *F1-measure*. From Table II, we can see that the SIFT-matching-based local stabilization is effective in improving the *Precision* score by reducing false detected foreground regions compared with the one without local stabilization.

In Fig. 6, we show foreground object detection results using one video as an example, where the illumination changes over time. We can see that the proposed method achieves the best performance: all objects are detected with a few false positives caused by shadows. Furthermore, it is capable of capturing objects with infrequent motions (e.g., the two persons near the top-right corner in frames 16350 and 21200), which all the other methods except the mean-shift-based method [51] fail to detect. The mean-shift-based method [51] has shown to suffer from local motions, and there is a lag in its background modeling (e.g., the person near the top-right corner in all frames is a false positive).

In Fig. 7, we show another example using a more challenging video. The background of this video includes many trees, bushes, and grasses, which are waving all the time; and the soldiers in camouflage uniforms have similar appearance/texture as the background. We can see that the mean-shift-based method [51] totally fails for this video because of the frequent local motions in the background. The ViBe [13] and the GRASTA [15] can detect only the soldiers partially when they are moving (see frames 12450 and 20850), and perform even worse when the persons are staying static (see other four frames). Furthermore, the ViBe detects many moving background. In contrast, the proposed method obtains the best detection result: much larger part of foreground and less background are detected.

2) Video Clips: For further evaluating the performance of detecting the objects with infrequent motions, we select 20 short clips from the DARPA data set. Each clip has 1000 frames and contains objects with infrequent motions: the objects stay at some locations for a relatively long time within the clip. A quantitative validation is performed on

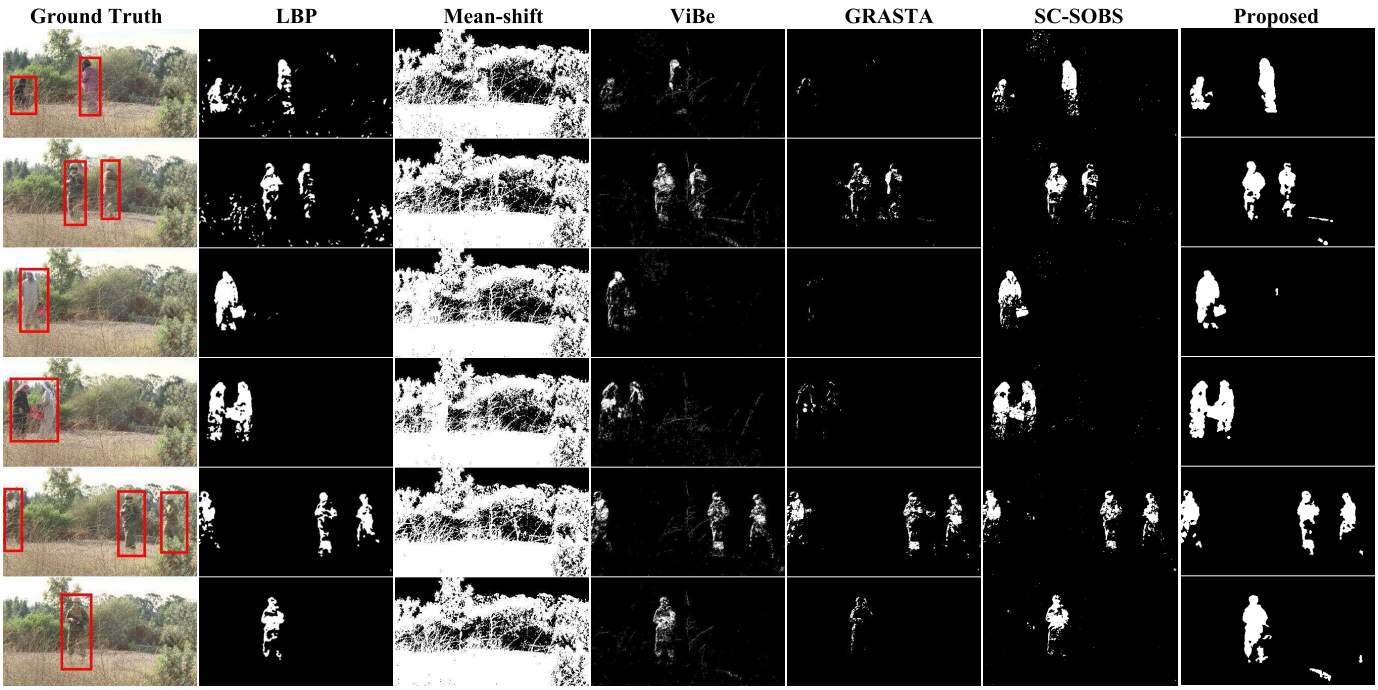


Fig. 7. Detection results on another sample video in the DARPA data set. Top to bottom: frames 4800, 12450, 15500, 17000, 20850, and 21100. Best viewed in color.

TABLE III
PERFORMANCE COMPARISON ON 20 VIDEO CLIPS, WHICH ARE SELECTED FROM THE DARPA DATA SET AND CONTAIN OBJECTS WITH INFREQUENT MOTIONS

Model	Recall	Precision	F1-measure
LBP based [36]	0.3198	0.6761	0.4342
Mean-shift based [51]	0.4338	0.4896	0.4600
ViBe [13]	0.4201	0.4802	0.4481
GRASTA [15]	0.2333	0.6808	0.3475
SC-SOBS [52]	0.4853	0.4611	0.4729
Proposed Forward	0.4752	0.6586	0.5521
Proposed (w/o stabilization)	**0.6021**	0.7710	0.6762
Proposed	0.6004	**0.7811**	**0.6789**

these 20 short clips and the experimental results are reported in Table III. In order to demonstrate the effectiveness of forward–backward background propagation, we also compare the proposed method with the one only with forward background propagation. As shown in Table III, the proposed visual-attention-based algorithms (the last three rows in Table III) including the one without the backward propagation outperform the other state-of-the-art methods in terms of *F1-measure*. Not surprisingly, the proposed method with forward-backward propagation and local motion stabilization yields the best foreground detection performance among all methods in comparison.

As shown in Fig. 8, a qualitative comparison is performed on one of the video clips. In this clip, a red bag is abandoned on the ground at the beginning of the clip and then is taken away. Most of the methods (the mean-shift based [51], the ViBe [13], and the GRASTA [15]) have a lag in their background modeling and hence produce false positive detection of the bag after it is taken away. The proposed methods (the last two columns in Fig. 8), even without backward propagation, can successfully detect the removed bag by employing the visual-attention analysis in the background modeling. Furthermore, with the forward-backward background propagation, the proposed method (the last column in Fig. 8) is able to detect the bag in the whole video.

3) Discussion on Forward/Backward Background Propagation: Most of the time, the results of using the forward–backward background propagation are comparable to the one using only the forward propagation, as shown in the first three columns of Fig. 9. For the applications, which require online processing, e.g., video surveillance, the proposed visual-attention-based method could work well in an online manner by only using the forward background propagation.

However, the forward propagation can only propagate the background along the time and thus, will fail to detect the foreground before it moves, i.e., the removed objects. As shown in the last three columns of Fig. 9, it fails to detect the people and the cart at the beginning of the video by using only the forward background propagation. Hence, for the applications in which online processing is not necessary, such as video retrieval, the forward–backward background propagation will enable a more robust background model, which is especially capable of handling foreground objects with infrequent motions.

B. Experimental Results on the ChangeDetection Data Set

In the second set of experiments, the proposed method has been evaluated on the ChangeDetection benchmark data set [5], [6][6] for the category of *intermittent object motion*.

[6]The workshop held in conjunction with IEEE Conference on Computer Vision and Pattern Recognition-2012 and 2014.

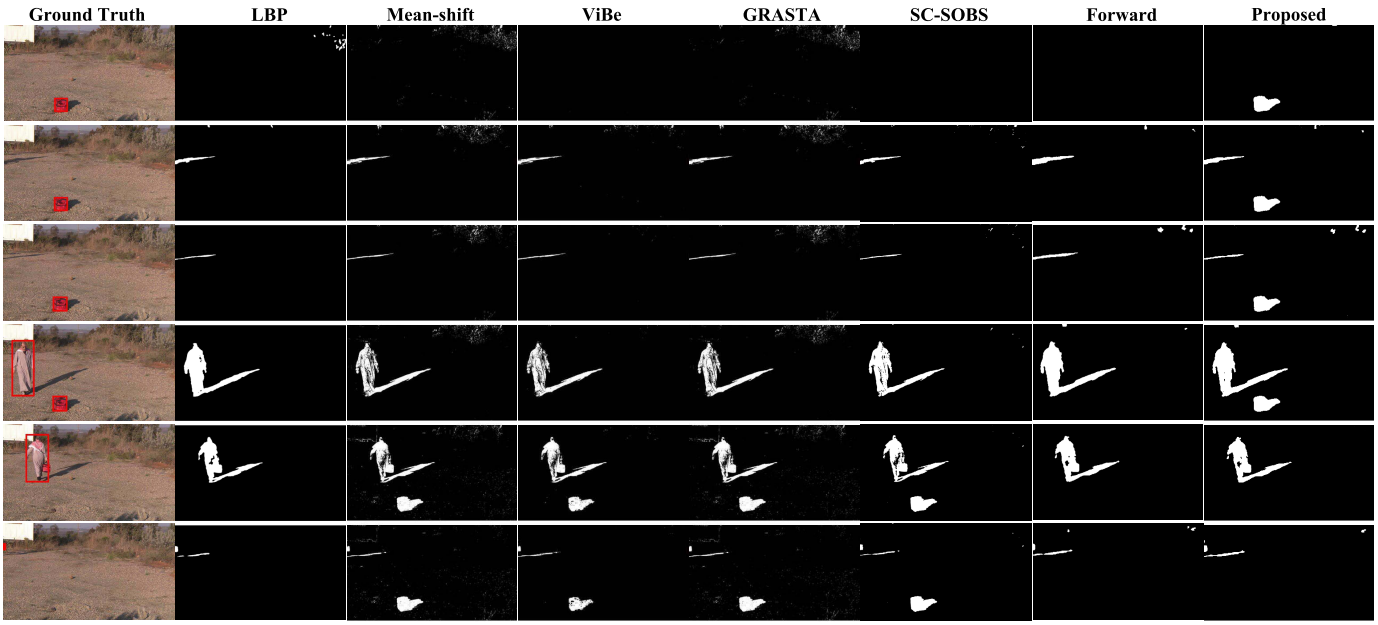


Fig. 8. Detection results on one of the selected video clips in the DARPA data set. The red duffle bag stays still at the beginning of the clip and is taken away. Top to bottom: frames 1, 200, 400, 600, 800, and 1000. We give the results only using the forward background propagation in the last column, which we will discuss in Section V-A3). Best viewed in color.

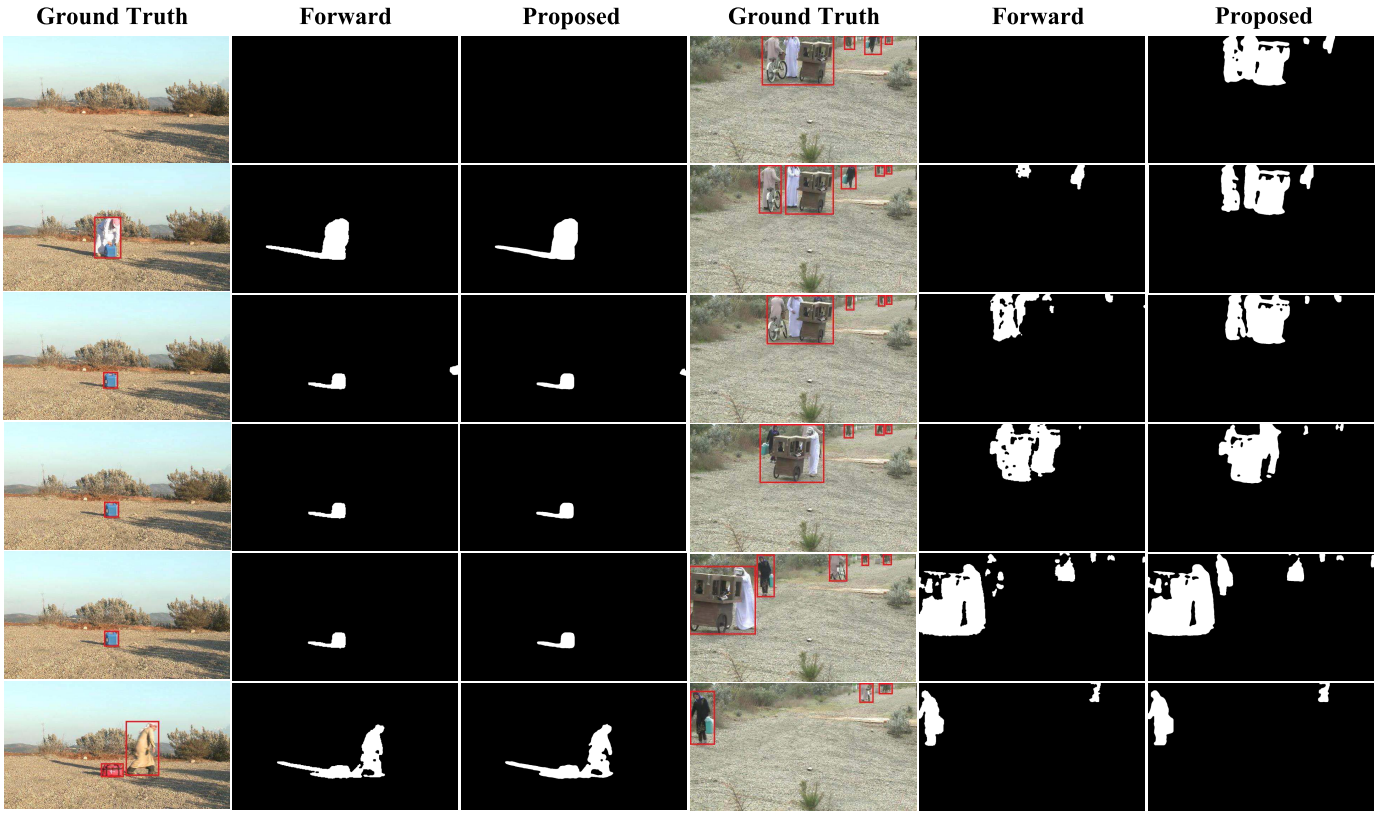


Fig. 9. Qualitative comparisons between the results of using only the forward propagation and the ones of using both forward and backward background propagation (proposed) on a DARPA video clip. Forward denotes the results of using only forward background propagation. Proposed denotes the results of using both the forward and backward background propagation. We show two cases in the first and last three columns, respectively. Top to bottom: frames 1, 200, 400, 600, 800, and 1000. Best viewed in color.

There are six videos (18 650 frames in total) in the *intermittent object motion* category, each of which contains objects with infrequent motions, e.g., abandoned objects and parked cars moving away. These objects are often treated as part of background before moving and will introduce ghost artifacts in background subtraction.

TABLE IV

PERFORMANCE COMPARISON ON THE INTERMITTENT OBJECT MOTION SUBDATA SET OF CHANGEDETECTION. IT IS WORTH MENTIONING THAT, BASED ON THE BENCHMARK'S WEBSITE, THE REPORTED BEST *F1-Measure* IS 0.7891 FROM [28]

Model	Recall	Specificity	FPR	FNR	PWC	Precision	F1-measure
LBP based [36]	0.6556	**0.9978**	**0.0022**	0.0353	3.2250	**0.9083**	0.7379
Mean-shift based [51]	**0.8709**	0.6213	0.3787	**0.0074**	35.3842	0.3223	0.3656
ViBe [13]	0.6811	0.9410	0.0590	0.0295	7.6330	0.5938	0.5847
GRASTA [15]	0.2135	0.9926	0.0074	0.0617	6.1645	0.5661	0.2768
SC-SOBS [52]	0.7237	0.9613	0.0387	0.2763	5.2207	0.5896	0.5918
Proposed	0.8241	0.9915	0.0085	0.0099	**1.6610**	0.8590	**0.8399**

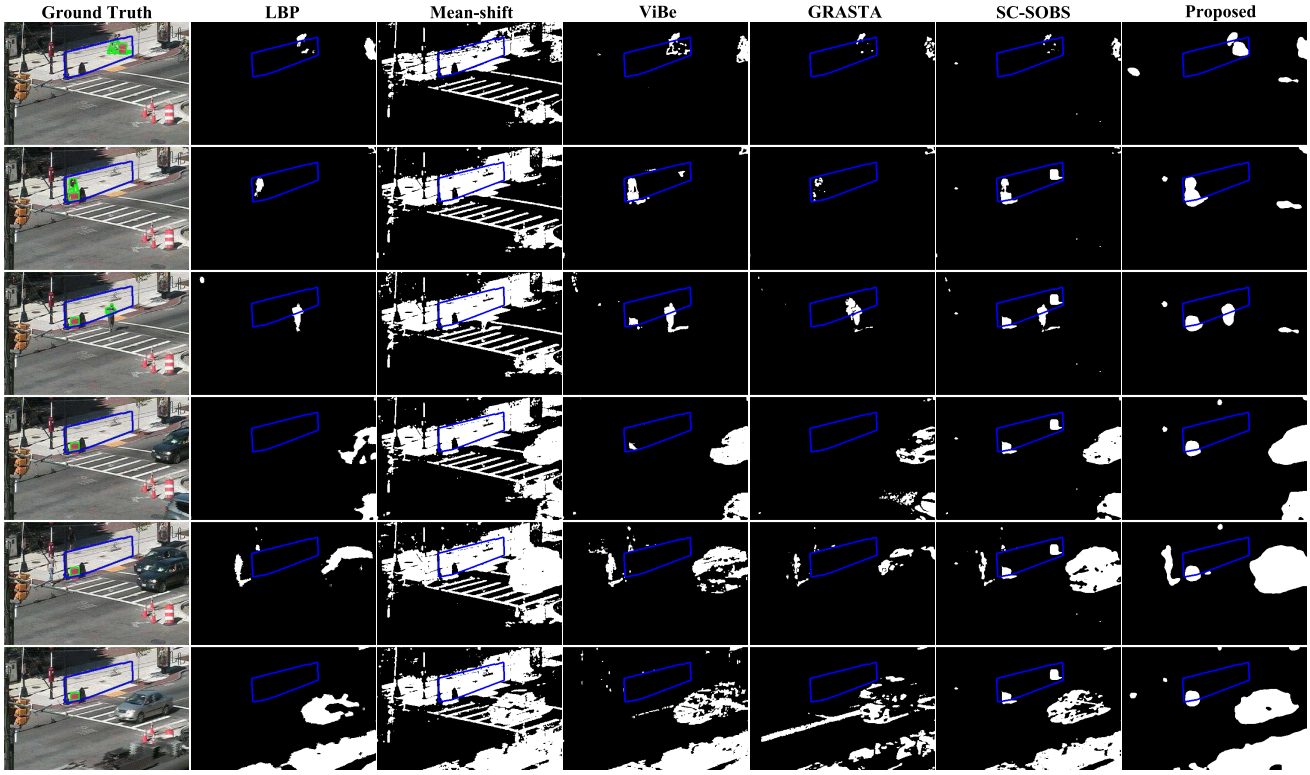


Fig. 10. Qualitative comparison on a video in ChangeDetection data set for the *intermittent object motion* category. The region inside the blue boundary is the ROI, where ground truth labels of foreground objects are provided. Note that a red box was moved from the right to the left of the ROI and stayed still for a long time, which is detected only by the proposed method. Top to bottom: frames 2450, 2900, 3300, 3700, 4100, and 4500. Best viewed in color.

The proposed method and other methods are quantitatively evaluated on the *intermittent object motion* category exactly following the benchmark evaluation procedure as in [5] and [6].[7] Specifically, by defining TP, TN, FN, and FP as the number of true positives, true negatives, false negatives, and false positives, respectively, four additional evaluation metrics are defined and employed in the benchmark: 1) *specificity:* $(\mathrm{TN}/\mathrm{TN}+\mathrm{FP})$; 2) *false positive rate (FPR):* $(\mathrm{FP}/\mathrm{FP}+\mathrm{TN})$; 3) *false negative rate (FNR):* $(\mathrm{FN}/\mathrm{TN}+\mathrm{FP})$; and 4) *percentage of wrong classifications (PWC):* $100\times(\mathrm{FN}+\mathrm{FP}/\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN})$.

As shown in Table IV, the proposed method outperforms all the comparison methods in terms of the *F1-measure* and the *PWC*, which consider both *Precision* and *Recall*. Furthermore, the proposed method achieved the best performance among all the other methods that evaluated their performance on the intermittent object motion subdata set of the ChangeDetection database, according to their results reported on the benchmark's website.[8]

We also present a qualitative comparison on the ChangeDetection data set, as shown in Fig. 10. The ground truth labels of foreground objects are provided only in the region of interest (ROI) denoted as the region enclosed by the blue boundary. From Fig. 10, we can see that the mean-shift-based method [51] generally produces more false positive detection of foreground objects than the other methods. Note that a red box was moved from the right to the left of the ROI and stayed still for a long time. The proposed method is able to capture this box all the time, while all the other methods in comparison fail to detect it. Furthermore, even for the regions outside the ROI, which are not counted in the evaluation, our method can detect the moving objects better with less false

[7]We directly run the evaluation code provided [5], [6].

[8]http://www.changedetection.net/

TABLE V
AVERAGE RUNNING TIME PER FRAME FOR THE FOUR MAJOR STEPS, i.e., INIT. (BACKGROUND INITIALIZATION), PROP. (BACKGROUND PROPAGATION), FG. (FOREGROUND DETECTION), AND STAB. (LOCAL MOTION STABILIZATION), AS WELL AS THE TOTAL TIME

Dataset	Frame Size	Init.	Prop.	Fg.	Stab.	Total
DARPA	1280×720	1.17ms	0.083s	0.22s	0.18s	0.48s
Change	432×288	0.33ms	0.049s	0.11s	0.12	0.28s
	320×240	0.25ms	0.033s	0.06s	0.08s	0.17s

positives, especially in frames 4100 and 4500. In addition, compared with the other methods except the mean-shift-based method, the proposed one can catch the foreground objects as a whole (for example, the vehicles in frames 3700, 4100, and 4500), which is desired for next level tracking and recognition tasks.

C. Algorithm Efficiency

The proposed method was implemented in MATLAB and evaluated on a PC with Intel Xeon W3565 CPU and 4-GB RAM. The average running time per frame is reported in Table V for both data sets. Specifically, we report the per-frame running time for all four major steps, i.e., background initialization, background propagation, foreground detection, and local motion stabilization, as well as the total running time per frame. Since the background initialization and background propagation are performed only on key frames, each of which represents a 60-frame video clip, we divide the per-key-frame running time of these two steps by 60 to compute their per-frame running time.

VI. CONCLUSION

In this paper, we proposed a novel method to detect moving foreground objects, which is especially capable of detecting objects with infrequent motions. Specifically, we improved the background subtraction method by integrating a visual-attention mechanism to distinguish the foreground and background. The identified background regions can be propagated back-and-forth along the whole super-clip. Furthermore, we also proposed an SIFT-matching-based local motion stabilization algorithm to deal with the frequent local motions in the scene. Extensive experimental validations on two challenging data sets have demonstrated that the proposed method outperforms the state-of-the-art background subtraction methods in comparison. As shown in the experimental results, the performance improvement is more impressive for detecting objects with infrequent motions.

In this paper, a simple video decomposition strategy has been used to divide the long video into super-clips and works well under the assumption that the camera keeps static in most of the time. In order to handle complicated camera motions, in the future, we plan to try more sophisticated video decomposition methods, such as [53], to generate super-clips.

The proposed bidirection background propagation strategy is suitable to build a background model in an offline manner. As we discussed in Section V-A3, the results of using only the forward propagation and using bidirection propagation are comparable, except for the removed objects. In the future, we plan to extend this paper to automatically switching between online and offline modes. The online mode, which uses only the forward propagation, is employed as the major mechanism for real-time foreground detection. Once a region that was previously modeled as part of background starts to move, i.e., an object is removed from the scene, the offline mode will be triggered and the background model will be updated using the bidirection background propagation. We also plan to use the proposed model in surveillance applications, especially when the events of interest involve infrequent moving objects, e.g., abandoned object detection and fall detection.

REFERENCES

[1] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1–9.

[2] X. Liu, L. Lin, S. Yan, and H. Jin, "Adaptive object tracking by learning hybrid template online," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 11, pp. 1588–1599, Nov. 2011.

[3] A. Shimad, S. Yoshinaga, and R.-I. Taniguchi, "Adaptive background modeling for paused object regions," in *Proc. Asian Conf. Comput. Vis. Workshops*, 2011, pp. 12–22.

[4] Y. Tian, R. S. Feris, H. Liu, A. Hampapur, and M.-T. Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 5, pp. 565–576, Sep. 2011.

[5] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. CVPR Workshops*, Jun. 2012, pp. 1–8.

[6] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. CVPR Workshops*, Jun. 2014, pp. 393–400.

[7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999.

[8] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 255–261.

[9] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[10] Z. Kim, "Real time object tracking based on dynamic feature grouping with background subtraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[11] T. Ko, S. Soatto, and D. Estrin, "Warping background subtraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1331–1338.

[12] J. K. Suhr, H. G. Jung, G. Li, and J. Kim, "Mixture of Gaussians-based background subtraction for Bayer-pattern image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 365–370, Mar. 2011.

[13] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[14] D. Sun, E. B. Sudderth, and M. J. Black, "Layered segmentation and optical flow estimation over time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1768–1775.

[15] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1568–1575.

[16] L. Cheng, M. Gong, D. Schuurmans, and T. Caelli, "Real-time discriminative background subtraction," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1401–1414, May 2011.

[17] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1937–1944.

[18] T. Bouwmans, "Recent advanced statistical background modeling for foreground detection: A systematic survey," *Recent Patents Comput. Sci.*, vol. 4, no. 3, pp. 147–176, 2011.

[19] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vis. Image Understand.*, vol. 122, pp. 4–21, May 2014.
[20] T. S. F. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, Apr. 2014.
[21] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. 6th Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.
[22] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.
[23] M. Narayana, A. Hanson, and E. Learned-Miller, "Background modeling using adaptive pixelwise kernel variances in a hybrid feature space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2104–2111.
[24] Y. Moshe, H. Hel-Or, and Y. Hel-Or, "Foreground detection using spatiotemporal projection kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3210–3217.
[25] F. J. Hernandez-Lopez and M. Rivera, "Change detection by probabilistic segmentation from monocular view," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1175–1195, 2014.
[26] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Comput. Soc. Conf. CVPR Workshops*, Jun. 2012, pp. 38–43.
[27] A. Shimada, H. Nagahara, and R.-I. Taniguchi, "Background modeling based on bidirectional analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1979–1986.
[28] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split Gaussian models," in *Proc. IEEE Conf. CVPR Workshops*, Jun. 2014, pp. 420–424.
[29] B. Wang and P. Dudek, "A fast self-tuning background subtraction algorithm," in *Proc. IEEE Conf. CVPR Workshops*, Jun. 2014, pp. 401–404.
[30] X. Cui, J. Huang, S. Zhang, and D. N. Metaxas, "Background subtraction using low rank and group sparsity constraints," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 612–625.
[31] L. Lin, Y. Xu, X. Liang, and J. Lai, "Complex background subtraction by pursuing dynamic spatio-temporal models," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3191–3202, Jul. 2014.
[32] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1301–1306.
[33] H. Han, J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Moving object detection revisited: Speed and robustness," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 910–921, Jun. 2015.
[34] N. Liu, H. Wu, and L. Lin, "Hierarchical ensemble of background models for PTZ-based video surveillance," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 89–102, Jan. 2015.
[35] S. Kim, D. W. Yang, and H. W. Park, "A disparity-based adaptive multi-homography method for moving target detection based on global motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[36] J. Yao and J.-M. Odobez, "Multi-layer background subtraction based on color and texture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[37] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Left-luggage detection from finite-state-machine analysis in static-camera videos," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4600–4605.
[38] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1359–1370, Jul. 2015.
[39] L. Maddalena and A. Petrosino, "Stopped object detection by learning foreground model in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 723–735, May 2013.
[40] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 853–860.
[41] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[42] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.
[43] E. Rahtu and J. Heikkilä, "A simple and efficient saliency detector for background subtraction," in *Proc. IEEE 12th ICCV Workshop*, Sep./Oct. 2009, pp. 1137–1144.
[44] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–6.
[45] Y. Lin, B. Fang, and Y. Tang, "A computational model for saliency maps by using local entropy," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 967–973.
[46] Z. Li, "A saliency map in primary visual cortex," *Trends Cognit. Sci.*, vol. 6, no. 1, pp. 9–16, Jan. 2002.
[47] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[48] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 314–328, Feb. 2013.
[49] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[51] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 106–129, 2010.
[52] L. Maddalena and A. Petrosino, "The SOBS algorithm: What are the limits?" in *Proc. IEEE Comput. Soc. CVPR Workshops*, Jun. 2012, pp. 21–26.
[53] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 540–555.

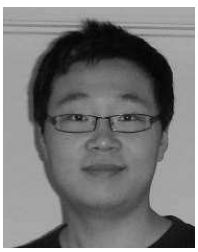
Yuewei Lin (S'09) received the B.S. degree in optical information science and technology from Sichuan University, Chengdu, China, in 2004 and the M.E. degree in optical engineering from Chongqing University, Chongqing, China, in 2007. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA.

His research interests include computer vision, machine learning, and image/video processing.

Yan Tong (M'12) received the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2007.

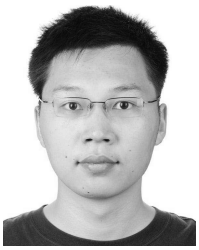
She was a Research Scientist with the Visualization and Computer Vision Laboratory, GE Global Research, Niskayuna, NY, USA, from 2008 to 2010. She is currently an Assistant Professor with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA. Her research interests include computer vision, machine learning, and human–computer interaction.

Dr. Tong has served as a Conference Organizer, an Area Chair, and a Program Committee Member for a number of premier international conferences.



Yu Cao (S'09–M'13) received the B.S. degree in information and computation science and the M.S. degree in applied mathematics from Northeastern University, Shenyang, China, in 2003 and 2007, respectively, and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2013.

He is currently a Software Engineering Researcher with the IBM Almaden Research Center, San Jose, CA, USA. His research interests include computer vision, machine learning, pattern recognition, and medical image processing.



Youjie Zhou (S'13) received the B.S. degree in software engineering from East China Normal University (ECNU), Shanghai, China, in 2010. He is currently pursuing the Ph.D. degree in computer science and engineering with the University of South Carolina, Columbia, SC, USA.

From 2007 to 2010, he was a Research Assistant with the Institute of Massive Computing, ECNU, where he was involved in multimedia news exploration and retrieval. He is a Research Assistant with the Computer Vision Laboratory, University of South Carolina. His research interests include computer vision, machine learning, and large-scale multimedia analysis.

Song Wang (SM'12) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign (UIUC), Urbana, IL, USA, in 2002.

From 1998 to 2002, he was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His research interests include computer vision, medical image processing, and machine learning.

Dr. Wang is a member of the IEEE Computer Society. He currently serves as the Publicity/Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society, and an Associate Editor of *Pattern Recognition Letters*.