Spatiotemporal GMM for Background Subtraction with Superpixel Hierarchy

Mingliang Chen, Xing Wei, Qingxiong Yang, Qing Li, Gang Wang, and Ming-Hsuan Yang

Abstract—We propose a background subtraction algorithm using hierarchical superpixel segmentation, spanning trees and optical flow. First, we generate superpixel segmentation trees using a number of Gaussian Mixture Models (GMMs) by treating each GMM as one vertex to construct spanning trees. Next, we use the M-smoother to enhance the spatial consistency on the spanning trees and estimate optical flow to extend the M-smoother to the temporal domain. Experimental results on synthetic and real-world benchmark datasets show that the proposed algorithm performs favorably for background subtraction in videos against the state-of-the-art methods in spite of frequent and sudden changes of pixel values.

Index Terms—Background modeling, superpixel hierarchy, minimum spanning tree, tracking, optical flow

1 Introduction

Background modeling is one of the most extensively studied topics in computer vision [1], [2], [3], [4]. It is usually used as a preprocessing step in numerous vision applications including video surveillance, event detection, and human-computer interface, to name a few. The increasing use of mobile phones has motivated the development of background subtraction methods in moving cameras, and recent methods [5], [6] using motion estimation for compensating the camera motion have demonstrated its effectiveness in background subtraction. However, the application domains are limited to rather strict assumptions such as low scene complexity. It remains a challenging problem to develop efficient and robust background subtraction algorithms, with the assumption of static cameras, to account for dynamic background, lighting changes and cluttered scenes. In this paper, we propose a real-time background subtraction algorithm using spatiotemporal cues from videos and demonstrate its effectiveness against the state-of-the-art methods on benchmark datasets.

1.1 Related Work

Background subtraction algorithms can be broadly categorized based on pixels [7], [8], [9], [10], [11], [12], [13], [14], block features [15], [16], [17], [18], [19], [20], regions [21], [22], [23], [24], clustering [25], [26], superpixels [27], [28], [29], [30], and hybrid cues [31], [32], [33].

Pixel based methods model pixel appearance by parametric probability density functions such as a mixture of Gaussians [8], or non-parametric approaches such as kernel density estimation functions [7], [9] and histograms of the historical pixel values [13]. While these methods and variants [10], [11], [12], [14] have been shown to be able to distinguish foreground and background pixels efficiently, they are sensitive to inevitable irregular background changes such as sudden illumination changes and camera jitter. Block features such as Local Binary Patterns (LBP) [16], [17], [18], [20] are developed based on local textures around a pixel to alleviate the effects of variant illuminations. However, they are less robust to frequent appearance changes of pixels. Region based methods consider spatial correlation to refine the raw pixel level classification and alleviate foreground aperture using region level background models [22] or foreground shape models [24]. Clustering based methods [25], [26] subtract the background with cluster density estimation to cope with slight movements in the background. Hybrid cues based approaches [31], [32], [33] have been proposed to combine the advantages of various cues in which spatial texture and temporal motion cues are used. In this work, we propose an algorithm to synergistically integrate efficient pixel based modality by using GMMs with spatiotemporal cues for robust background subtraction.

The spatiotemporal constraints for GMMs have been shown as robust to sudden changes in [34]. In [35], a compact representation of texture and motion patterns in each block of the video frame is developed to account for appearance changes caused by background motion. However, the pixels near the block boundaries are not modeled well. Fang et al. [36] use color pixels and surrounding neighbors features to construct the GMMs to detect objects more effectively at the cost of heavy computational loads. Markov Random Filed (MRF) are widely used in background modeling to enforce the spatial and temporal contiguity [37], [38], [39]. However, the object boundaries are usually less accurate, or the computational cost is high. For efficiency and effectiveness, we integrate the Minimum Spanning Tree (MST) based aggregation method [40] with a robust estimator for a spatially-consistent solution, and enforce temporally-consistent constraints with a fast edge-preserving optical flow algorithm [41].

Recently superpixels have been exploited in video object segmentation methods for increasing spatial coherency. Both appearance and motion models for each superpixel are used to determine labels for each pixel with belief propagation [27]. In [28], it first generates coarse foreground segmentations that predict motion regions by analyzing how superpixels change in consecutive frames; the segmentations are next refined based on appearance and perceptual organization on motion regions. A superpixel-based matrix decomposition method [29] is developed to exploit sparsity and structured foreground constraints for efficient background subtraction. Nevertheless, methods using one single layer of superpixels are still not effective for foreground objects undergoing large scale changes or background with dynamic appearance variations. Thus, a background subtraction algorithm based on a superpixel hierarchy is proposed in this work. Different from the method [30] which simply captures foregrounds under different scales and averages the multi-scale segmentations, we use hierarchical GMMs for the background model to handle large scale and dynamic appearance changes. Through integrating with spatiotemporal cues of superpixels at each scale, we show the proposed algorithm performs favorably in complex scenes at low computational cost.

1.2 Context and Contributions

We propose an algorithm to exploit a superpixel hierarchy for spatiotemporally-consistent background subtraction based on our earlier work [42]. The GMM [8] is used to construct an initial background estimate at each individual pixel location. An efficient

- *M. Chen and Q. Li are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. E-mail: christmas.chen.34@gmail.com, itqli@cityu.edu.hk.*
- *Q. Yang is with the School of Information Science and Technology, University of Science and Technology of China, Anhui 230000, China. E-mail: liiton.research@gmail.com.*
- *X. Wei is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Shaanxi 710049, China. E-mail: xingxjtu@gmail.com.*
- *G. Wang is with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798. E-mail: wanggang@ntu.edu.sg.*
- *M.-H. Yang is with the School of Engineering, University of California, Merced, CA 95340. E-mail: mhyang@ucmerced.edu.*

Manuscript received 26 May 2016; revised 27 Apr. 2017; accepted 7 June 2017. Date of publication 20 June 2017; date of current version 14 May 2018.
(Corresponding author: Ming-Hsuan Yang.)
Recommended for acceptance by R. Collins.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2717828

MST based aggregation method [40] is integrated with an M-smoother to refine the initial estimates for a spatially-consistent solution. The GMMs with spatial constraint are thus robust to both frequent and sudden changes of pixel values. Compared to the original GMM [8], the additional computational cost is the MST based M-smoother, which can be obtained efficiently. Optical flow estimation is used to extend the proposed MST based M-smoother to enhance temporal consistency. Since appearance changes almost always exist in the background regions especially in the outdoor environments. As a result, we develop a background model based on a superpixel hierarchy to cope with the noise due to small motions of background. The main differences between this work and our earlier results [42] are summarized as follows:

1) *Superpixel hierarchy*. We propose a background model based on a superpixel hierarchy, from which the GMMs in different hierarchies together determine the background probability for each pixel.
2) *Spatiotemporal model with a hierarchical structure*. Enforcing spatiotemporal constraints on superpixels at each scale, our hierarchical GMMs can be more effective to account for appearance changes in dynamic background scenarios.
3) *Extensive performance evaluation*. Experimental results on both synthetic pixel-level SABS [43] and real-world region-level ChangeDetection [44] datasets demonstrate that our proposed algorithm performs favorably against the state-of-the-art methods.

2 Background Subtraction via Minimum Spanning Tree and Optical Flow

We briefly review the Gaussian mixture background model and present the pixel-based Spatially-consistent Background Model (SBM). We then propose the pixel-based SpatioTemporally-consistent Background Model (STBM) for videos.

2.1 Gaussian Mixture Background Model

Stauffer and Grimson [8] model the intensity value of a pixel by a mixture of Gaussians for background estimation. A pixel is considered to be background only when at least one Gaussians model includes its pixel value with sufficient and consistent evidence. The probability of observing a pixel value I_p^t at pixel p for frame t is represented by

$$P(I_p^t) = \sum_{k=1}^{K} w_k^t \cdot \eta\Big(I_p^t, \mu_k^t, \Sigma_k^t\Big), \tag{1}$$

where η is a Gaussian probability density function

$$\eta\Big(I_p^t, \mu_k^t, \Sigma_k^t\Big) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_k^t|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Big(I_p^t - \mu_k^t\Big)^{\top}\Big(\Sigma_k^t\Big)^{-1}\Big(I_p^t - \mu_k^t\Big)\right), \tag{2}$$

and μ_k^t as well as Σ_k^t is the mean value and the covariance matrix of the kth Gaussian at time t, respectively. Each pixel is described by K different Gaussian distributions. To adapt to illumination changes, the pixel values from the current frame are used to update the mixture model as long as they can be represented by a Gaussian model.

2.2 Spatially-Consistent Background Modeling

Each pixel in GMM is processed independently and thus less robust to noise. To better account for both sudden and frequent intensity changes, we incorporate region cues in the proposed spatially-consistent background model.

2.2.1 Using Minimum Spanning Tree

We assume that connected pixels with similar pixel values have similar background estimates, and thus spatially-consistent background subtraction can be obtained. The similarity between every two pixels is defined based on the minimax path [45] by treating a video frame as a undirected graph $G = (V, E)$. The vertices V are all image pixels, and the edges E are all the edges between the nearest neighboring pixels. Each minimax path identifies a region boundary without high contrast and does not cross the boundary of any thin-structured homogeneous object. Furthermore, each minimax path can be efficiently extracted using a minimum spanning tree [46], e.g., the method [47] which has linear time complexity in the number of pixels.

Let $d(p, q)$ denote the minimax path between a pair of nodes $\{p, q\}$ in the current frame I^t, and $b_p^t = \{0, 1\}$ denote the corresponding binary background estimates at pixel p obtained from a Gaussian mixture background model. The minimax path $d(p, q)$ (which is symmetric) is then filtered with an M-smoother [48] to handle outliers in the coarse estimates from a mixture of Gaussians. The refined background estimate is

$$b_p^{t,spatial} = \arg\min_i \sum_{q \in I^t} \exp\left(-\frac{d(p,q)}{\sigma}\right)|i - b_q^t|^{\alpha}. \tag{3}$$

When $\alpha = 1$, (3) is a weighted median filter that utilizes the minimax path length based on the underlying regularity of the video frame. Since $b_p^t = \{0, 1\}$, we have

$$b_p^{t,spatial} = \begin{cases} 1 & \text{if } \sum_{q \in I^t} \exp\left(-\frac{d(p,q)}{\sigma}\right) \cdot b_q^t > \sum_{q \in I^t} \exp\left(-\frac{d(p,q)}{\sigma}\right) \cdot |1 - b_q^t|, \\ 0 & \text{else.} \end{cases} \tag{4}$$

Let $\mathcal{B}^t$ denote an image where the pixel value at pixel q is b_q^t and $\mathcal{F}^t$ denote an image where the pixel value at pixel q is $|1 - b_q^t|$ at time t. In addition, let

$$\mathcal{B}_p^{t,\downarrow} = \sum_{q \in I^t} \exp\left(-\frac{d(p,q)}{\sigma}\right)\mathcal{B}_q^t, \tag{5}$$

and

$$\mathcal{F}_p^{t,\downarrow} = \sum_{q \in I^t} \exp\left(-\frac{d(p,q)}{\sigma}\right)\mathcal{F}_q^t, \tag{6}$$

denote the weighted aggregation results of image $\mathcal{B}^t$ and $\mathcal{F}^t$. Thus, (4) becomes

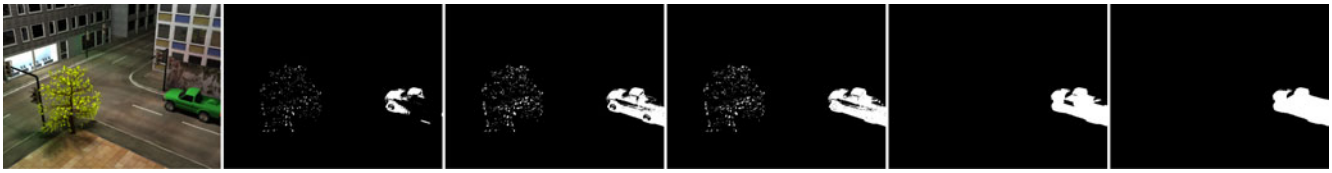
$$b_p^{t,spatial} = \begin{cases} 1 & \text{if } \mathcal{B}_q^{t,\downarrow} > \mathcal{F}_q^{t,\downarrow}, \\ 0 & \text{else.} \end{cases} \tag{7}$$

The background estimate $b_p^{t,spatial}$ obtained from the proposed MST-based M-smoother is used with the original estimate b_p^t to adjust K Gaussian distributions, and the only difference is that the distributions remain unchanged if either $b_p^{t,spatial}$ or b_p^t classifies pixel p as a foreground pixel. The effect from noisy background pixels on updating distributions can be significantly reduced using the spatially-consistent background estimates.

As shown in Fig. 1b, while part of the moving vehicle on the bottom right is continuously detected as the background using a GMM, the proposed MST-based M-smoother uses $b_p^{t,spatial}$ as constraints to update Gaussian distributions and better detect foreground pixels as shown in Fig. 1c.

2.2.2 Linear Time Solution

From (7), the main computational load of the proposed M-smoother lies in the weighted aggregation step in (5) as well as (6). As a brute-force implementation of the nonlocal aggregation



(a) Video Frame (b) GMM [8] (c) SBM (d) STBM (e) SBM with SH (f) STBM with SH

Fig. 1. Spatiotemporal background subtraction with a superpixel hierarchy. (a) a video frame extracted from the SABS dataset [43]. (b)-(f) foreground masks obtained from GMM, the proposed spatially-consistent and spatiotemporally-consistent background model without and with a superpixel hierarchy respectively.

step is computationally expensive, we use the recursive matching cost aggregation method [40],

$$\mathcal{B}_p^{t,\downarrow} = \exp\left(-\frac{d(P(p),p)}{\sigma}\right)\cdot \mathcal{B}_{P(p)}^{t,\downarrow} + \left(1-\exp\left(-\frac{2*d(p,P(p))}{\sigma}\right)\right)\cdot \mathcal{B}_p^{t,\uparrow}, \tag{8}$$

where $P(p)$ denotes the parent of node p, and

$$\mathcal{B}_p^{t,\uparrow} = \mathcal{B}_p^t + \sum_{P(q)=p} \exp\left(-\frac{d(p,q)}{\sigma}\right)\cdot \mathcal{B}_q^{t,\uparrow}. \tag{9}$$

Note that for 8-bit gray-scale images, $d(P(p),p)\in[0,255]$ and $d(p,q)\in[0,255]$ (when $P(q)=p$) and thus the values of $\exp(-\frac{d(P(p),p)}{\sigma})$ and $\exp(-\frac{d(p,q)}{\sigma})$ can be extracted from a single lookup table, and the value of $\left(1-\exp(-\frac{2*d(p,P(p))}{\sigma})\right)$ can be extracted from another table. Let T_1 and T_2 denote the two lookup tables, (8) and (9) can be written as

$$\mathcal{B}_p^{t,\downarrow} = T_1[d(P(p),p)]\cdot \mathcal{B}_{P(p)}^{t,\downarrow} + T_2[d(p,P(p))]\cdot \mathcal{B}_p^{t,\uparrow}, \tag{10}$$

$$\mathcal{B}_p^{t,\uparrow} = \mathcal{B}_p^t + \sum_{P(q)=p} T_1[d(p,q)]\cdot \mathcal{B}_q^{t,\uparrow}. \tag{11}$$

As only two additions and three multiplications are required at each pixel location, the proposed algorithm is computationally efficient.

2.3 Temporally-Consistent Background Modeling

2.3.1 Using Optical Flow

We extend the spatially-weighted M-smoother for background subtraction to the temporal domain

$$b_p^{t,temporal} = \arg\min_i \sum_{j=1}^{t}\sum_{q_j\in I^j} W(p,q_j)|i-b_{q_j}^j|, \tag{12}$$

where the similarity measurement is defined by

$$W(p,q_j) = \begin{cases} 1 & \text{if } q_j \text{ is the correct correspondence of } p \text{ in frame } j,\\ 0 & \text{else.}\end{cases} \tag{13}$$

In (13), $W(p,q_j)$ is obtained directly from the optical flow with the assumption that the background estimates for the same object appearing in two frames should be identical. As a trade-off between accuracy and speed, we use the edge-preserving patch match method [41] in this work.

Let $\Delta_p^{t,j}$ denote the motion vector between pixel p in frame t and the corresponding pixel $p_j = p+\Delta_p^{t,j}$ in frame j, and

$$v_p^t = \sum_{j=1}^{t} |b^j_{p+\Delta_p^{t,j}}|. \tag{14}$$

We simplify (12) by

$$b_p^{t,temporal} = \arg\min_i \sum_{j=1}^{t} |i-b^j_{p+\Delta_p^{t,j}}|, \tag{15}$$

$$= \begin{cases} 1 & \text{if } v_p^t > \frac{t}{2},\\ 0 & \text{else.}\end{cases} \tag{16}$$

A straightforward implementation of (15) is computationally expensive as the optical flow is estimated between any two frames, and $\frac{t(t-1)}{2}$ image pairs need to be computed to obtain the motion vectors $\Delta_p^{t,j}$ for $j\in[1,t-1]$. In practice, a recursive implementation is used to approximate v_p^t in (14) such that optical flow estimation is required only between every two successive frames

$$v_p^t = v^{t-1}_{p+\Delta_p^{t,t-1}} + |b_p^t|. \tag{17}$$

2.3.2 Spatiotemporal Background Model

A spatiotemporally-consistent background model can be directly obtained from (15) by replacing b_p^t with the spatially-consistent background estimates $b_p^{t,spatial}$ in (17)

$$v_p^t = v^{t-1}_{p+\Delta_p^{t,t-1}} + |b_p^{t,spatial}|. \tag{18}$$

3 Background Subtraction Using Superpixel Hierarchy

We propose an efficient spatiotemporal background subtraction algorithm based on hierarchical superpixel segmentations to handle inevitable background motion. We first generate a tree using a superpixel hierarchy. To integrate with the background model described in Section 2, the segmented superpixels are organized by a spanning tree. As the number of superpixels in a tree can be arbitrary (from one to image size), the computational complexity of an efficient segmentation method should be independent of the number of superpixels.

In this work, we use the superpixel hierarchy (SH) method [49], which has been shown to be computationally efficient. An input image is represented by a graph where the weights are dynamically adjusted to extract superpixels. Different from existing superpixel approaches that only generate a fixed number of superpixels at one time, superpixels of all scales can be generated concurrently, and those of the same scale conform to a tree topology. For example, an image of 480×320 pixels can be processed in 31 milliseconds to generate a superpixel hierarchy on a machine with a 2.3 GHz i7 CPU.

We extend the SH method [49] for background subtraction. First, we replace the intensity or color features with the Gaussian mixture model at each pixel. The GMMs serve as the vertices V in the graph $G=(V,E)$ while the edges E are all the edges between the nearest neighboring pixels. The edge cost is the Kullback-Leibler divergence computed from the two vertices (i.e., two GMMs) of the edge. Similar to [49], a superpixel hierarchy is

extracted from this graph based on the Borůvka's algorithm in linear time using edge contraction scheme [50].

After constructing an ordered spanning tree, any number of superpixels can be generated on the fly. The GMMs within a superpixel are merged into a single GMM (representing the corresponding superpixel) using the adaptive method [51] by varying the number of Gaussians. Thus, a spanning tree in which each vertex as a GMM is obtained.

A background estimate can be computed from each GMM/superpixel and then used in the background model presented in Section 2. Let $\mathcal{F}(\cdot)$ denote the background model presented in Section 2, and $b_{\mathbf{P}}^{init}$ and $b_{\mathbf{P}}^{final}$ denote the background estimate obtained from the GMM of superpixel $\mathbf{P}$. The background estimate obtained from $\mathcal{F}(\cdot)$ for superpixel $\mathbf{P}$ is

$$b_{\mathbf{P}}^{final} = \mathcal{F}(b_{\mathbf{P}}^{init}). \tag{19}$$

In practice, it is difficult to choose the optimal number of superpixels to be segmented in an image. As such, we propose to combine background estimates based on different numbers of superpixels. The maximum number of superpixels is the image size, and the number of superpixels is iteratively reduced by half until it is smaller than a threshold, e.g., 10 in this work. These superpixels result in a hierarchical segmentation tree, and each scale corresponds to a specific number of superpixels. Nevertheless, the first few scales (except for the first scale) is not used in order to maintain both the efficiency and accuracy.

Let C_p^s denote the mean color of the superpixel at sth scale containing pixel p, and $b_p^{s,init}$ as well as $b_p^{s,final}$ denote the input and output background estimates (at pixel p with sth scale) of the background model proposed in Section 2, respectively. In this case, $b_p^{s,init}$ and $b_p^{s,final}$ are the initial and final background estimates at pixel p at scale s, and according to (19)

$$b_p^{s,final} = \mathcal{F}(b_p^{s,init}). \tag{20}$$

Also let b_p^s denote the background estimate obtained directly from the GMM of the corresponding superpixel and assume that there are N scales (the one at scale N contains the maximum number of superpixels), we have

$$b_p^{N,init} = b_p^N, \tag{21}$$

$$\beta = \exp(-0.5 \parallel C_p^s, C_p^{s+1} \parallel^2 / \sigma_{sh}^2), \tag{22}$$

$$b_p^{s,init} = (1-\beta) \cdot b_p^s + \beta \cdot b_p^{s+1,final}, \tag{23}$$

where σ_{sh} is a constant (e.g., 10 in this work). The combined background estimate at pixel p is $b_p^{1,final}$. Figs. 1e and 1f show the background estimates based on a superpixel hierarchy by the proposed SBM and STBM. With the use of hierarchical superpixels, both models can robustly account for the backgrounds with dynamic appearance changes (e.g., waving trees) and generate better foreground segmentations.

Algorithm 1 summarizes the main steps of the proposed approach for background subtraction. Note that we also exploit spatial information and adopt a random strategy to effectively update the model to alleviate the issues with purely conservative schemes as suggested in [13].

4 Experimental Results

The proposed background subtraction algorithms are experimentally validated with a variety of scenes using two benchmark datasets. The Stuttgart Artificial Background Subtraction (SABS) [43] dataset consists of synthetic video frames, and the ChangeDetection 2012 [44] dataset contains real-world video frames. Both qualitative and quantitative evaluations with the state-of-the-art methods are presented. The methods based on hierarchical superpixels presented in Section 3 are referred to as the Spatially-consistent Superpixel Hierarchy Background Model (SSHBM) and SpatioTemporally-consistent Superpixel Hierarchy Background Model (STSHBM). More experimental results are available at http://www.cs.cityu.edu.hk/~mlchen2/publications/st_background/.

Algorithm 1. Spatiotemporal Background Subtraction

Initial:

$N = \lfloor log_2(\frac{pixel\ numbers}{10}) \rfloor$; // number of hierarchy

$GMM_N^1 \Leftarrow$ first frame;

$tc = 1$; $v^1[N] = 1$; // temporally-consistent indicator

Classify and Update:

for each new frame t **do**
  construct the MST $T_{spatial}^t$ with pixel color of t;
  use optical flow to estimate the motion vector $\Delta_p^{t,t-1}$ and the corresponding pixels $(p + \Delta_p^{t,t-1}, p)$;
  construct the second MST T_{gmm}^t with the GMM^{t-1} for obtaining the hierarchical superpixels $\mathbf{P}[N]$;
  $GMM_N^t = GMM_N^{t-1}$;
  for $s = N; s > 0; s--$ **do**
    $b_s^t \Leftarrow GMM_s^t(I^t)$;
    if $s == N$ **then**
      $b_s^{t,init} = b_s^t$;
    else
      $b_s^{t,init} = (1-\beta) \cdot b_s^t + \beta \cdot b_{s+1}^{t,final}$;
    end if
    $b_s^{t,spatial} \Leftarrow T_{spatial}^t(b_s^{t,init})$;
    $tc = tc + 1$; $v_p^t[s] = v_{p+\Delta_p^{t,t-1}}^{t-1}[s] + b_{p,s}^{t,spatial}$;
    $b_s^{t,temporal} = v^t[s] > \frac{tc}{2} ? 1 : 0$;
    $b_s^{t,final} = b_s^{t,temporal}$;
    $GMM_{s-1}^t \Leftarrow merge(GMM_s^t)$ if $s > 1$;
  end for
  // update model
  for each pixel p **do**
    if $b_{p,1}^{t,final} == 0$ **then**
      use I_p to update the GMM_p^t;
    else
      $q = getRandomNeighbour(p)$;
      use I_q to update the GMM_p^t if $b_{q,1}^{t,final} == 0$;
    end if
    // reset the indicator for the stationary pixel
    set $tc_p = 1$ and $v_p^t[H] = 1$ if $\Delta_p^{t,t-1} == (0,0)$
  end for
end for

4.1 Evaluation Metric and Parameter Settings

Each algorithm is evaluated based on the binary segmentation result at each pixel in terms of True Positive (TP), False Positive (FP) and False Negative (FN) rates using the ground-truth label. The F-measure is computed based on both precision P and recall R by $F = 2\frac{R \cdot P}{R+P}$, where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$. As in the reported results on the SABS dataset, the maximal F-measures (average over sequences) are used for evaluation, and the value of σ in Section 2.2 is accordingly set. For all the other experiments, the value of σ is set to 0.1.

On the other hand, the value of σ_{sh} in Section 3 is set to 10 to exploit the background estimates at different superpixel scales. The results based on this setting show the advantages of the SSHBM and STSHBM as well as how hierarchical superpixels can be used to improve background subtraction beyond raw pixels.

TABLE 1
F-Measures on the SABS Dataset [43]

Approach	*Basic*	*Dynamic Background*	*Bootstrap*	*Darkening*	*Light Switch*	*Noisy Night*	Average
McFarlane [52]	0.614	0.482	0.541	0.496	0.211	0.203	0.425
Stauffer [8]	0.800	0.704	0.642	0.404	0.217	0.194	0.494
Oliver [53]	0.635	0.552	-	0.300	0.198	0.213	0.380
McKenna [54]	0.522	0.415	0.301	0.484	0.306	0.098	0.354
Li [55]	0.766	0.641	0.678	0.704	0.316	0.047	0.525
Kim [56]	0.582	0.341	0.318	0.342	-	-	0.396
Zivkovic [11]	0.768	0.704	0.632	0.620	0.300	0.321	0.558
Maddalena [57]	0.766	0.715	0.495	0.663	0.213	0.263	0.519
Barnich [13]	0.761	0.711	0.685	0.678	0.268	0.271	0.562
AtsushiShimada [58]	0.723	0.623	0.708	0.577	0.335	0.475	0.574
Proposed SBM	0.764	0.747	0.669	0.672	0.364	0.519	0.623
Proposed STBM	0.813	0.788	0.736	0.753	0.515	0.680	0.714
Proposed SSHBM	**0.815**	**0.795**	**0.742**	**0.774**	**0.598**	**0.692**	**0.736**
Proposed STSHBM	**0.846**	**0.804**	**0.797**	**0.820**	**0.684**	**0.755**	**0.784**

The best two results are shown in red and blue.

4.2 Evaluation on the SABS Dataset

The SABS dataset contains six image sets with diverse scene changes designed for performance evaluation of background subtraction methods. The *dynamic background* set contains frequent or irregular movements in the background. As the *bootstrapping* set has no initialization images, the background subtraction task starts after the first frame. In the *darkening* set, the contrast between background and foreground is decreased by varying illumination gradually. The images of the *light switch* set are recorded with sudden illumination changes. In the *noisy night* set, the images are acquired with a significant amount of sensor noise. Each set contains 600 frames except the *darkening* and *bootstrapping* where each has 1,400 frames. All images of 800×600 pixels are captured at fixed viewpoints.

Table 1 shows the maximal F-measures of the proposed spatio-temporal background subtraction models (SBM, STBM, SSHBM, and STSHBM) and the reported results by the state-of-the-art methods. Overall, the proposed algorithms perform favorably against the others on this benchmark dataset. We note the proposed STBM performs well against the recent method with a bidirectional GMM [58] in every set. In addition, the extended SSHBM and STSHBM based on hierarchical superpixels outperform the pixel-based SBM and STBM for background subtraction.

The precision-recall curves with respect to different challenging factors are presented in Fig. 2. The proposed STSHBM achieves the highest recall ratios at the same precision levels with four challenging factors: dynamic background, gradual illumination changes (*darkening*), sudden illumination changes (*light switch*) and sensor noise (*noisy night*). Note that region-based background subtraction methods are less robust to the dynamic background while pixel-level background models are not robust to the sudden illumination changes and sensor noise. The proposed models exploit the properties of these two approaches to model background changes with the superpixel hierarchies.

4.3 Evaluation on the ChangeDetection Dataset

All the 31 video sequences in the ChangeDeteciton 2012 dataset with labeled ground truth are used for performance evaluation. Similar to the SABS dataset, the video sequences are categorized into six sets based on different challenging factors. The *dynamic background* set contains images of outdoor scenes with significant background motion. The *camera jitter* set contains videos captured by moving cameras, and the *shadows* set contains scenes with different levels of shadows. The *intermittent object motion* set contains videos with scenarios known for causing ghost artifacts in the background subtraction. The *thermal* set consists of videos captured by far-infrared cameras that contain significant artifacts.

Table 2 shows the evaluation results of the proposed algorithms on different subsets in terms of F-measure. While the proposed SBM and STBM perform well in most categories, the background models based on pixels are less effective to account for appearance changes in complex scenes when compared to the state-of-the-art approaches. The SSHBM and STSHBM with superpixel hierarchies perform significantly better especially in the *dynamic background*, *camera jitter*, *intermittent object motion* and *shadow* sets. Note that there exist significant appearance changes in the background regions for all these video sets due to object motion, camera

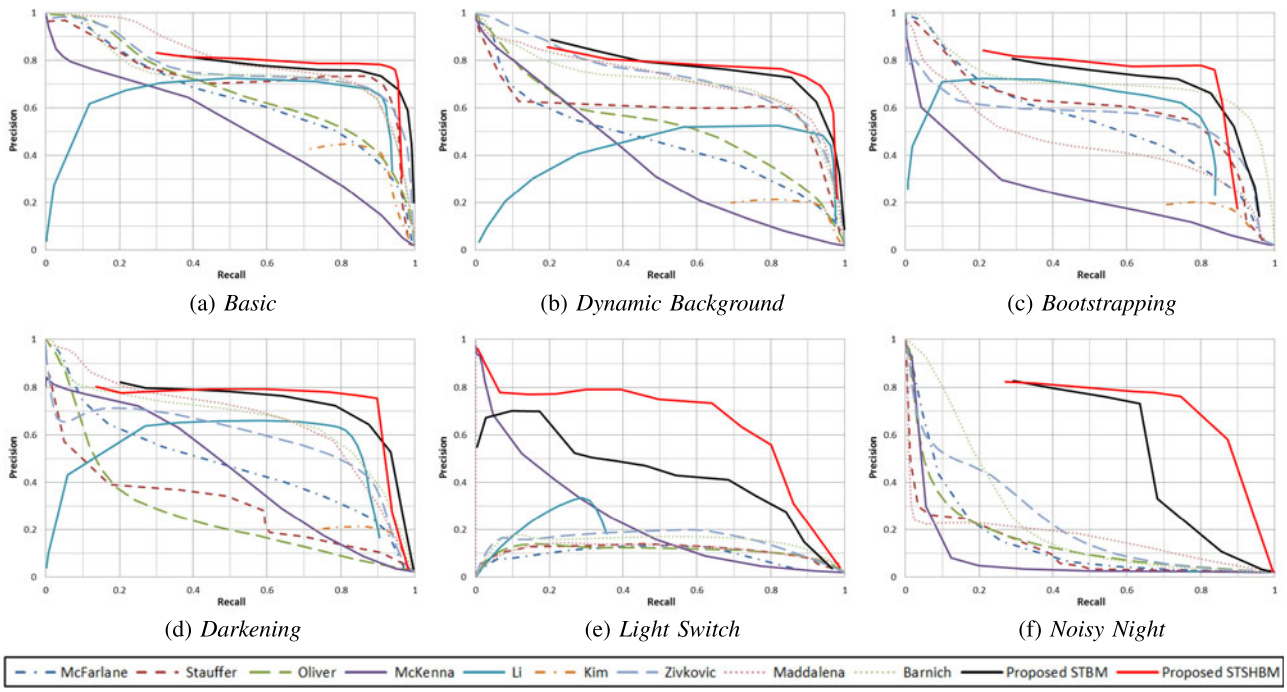


Fig. 2. Precision-recall curves on the SABS dataset [43] with different challenging factors. The red and dark solid curves show the performance of the proposed spatio-temporal background subtraction algorithms with and without superpixel hierarchy. Overall, the proposed algorithms (especially the STSHBM) perform favorably against all the other methods.

TABLE 2
F-Measures for the ChangeDetection 2012 Dataset

Approach	*Baseline*	*Dynamic Background*	*Camera Jitter*	*Intermittent Motion*	*Shadow*	*Thermal*	Average
Spectral-360 [59]	0.9330	0.7872	0.7156	0.5656	0.8843	0.7764	0.7770
CwisarD [60]	0.9075	0.8086	0.7814	0.5674	0.8412	0.7619	0.7780
GPRMF [61]	0.9280	0.7726	**0.8596**	0.4870	**0.8889**	0.8305	0.7944
SuBSENSE [62]	**0.9503**	0.8177	0.8152	0.6569	0.8646	0.8305	0.8260
PAWCS [63]	0.9397	0.8938	0.8137	0.7764	0.8710	0.8324	0.8579
Proposed SBM	0.9250	0.7882	0.7413	0.6755	0.8458	0.8423	0.8030
Proposed STBM	0.9345	0.8193	0.7522	0.6780	0.8529	**0.8571**	0.8157
Proposed SSHBM	0.9428	**0.9008**	0.8034	**0.8001**	0.8788	0.8443	**0.8617**
Proposed STSHBM	**0.9534**	**0.9120**	**0.8503**	**0.8349**	**0.8930**	**0.8579**	**0.8836**

The best two results are shown in red and blue.

TABLE 3
Computational Cost of the Proposed Background Subtraction Algorithms for QVGA Videos (Milliseconds/Frame)

Method	GMM [8]	BGMM [58]	Proposed					
			SBM	STBM		SSHBM	STSHBM	
	CPU	CPU	CPU	CPU	GPU	CPU	CPU	GPU
Time	12	**5**	15	982	83	54	1,020	90

movement, and lighting. Overall, the proposed STSHBM performs favorably against the other methods in all these categories. The SSHBM and STSHBM approaches outperform all the other algorithms on average, which can be attributed to the use of superpixel hierarchies and proposed spatiotemporal background model. Fig. 3 shows some sample results from the evaluated background subtraction methods.

4.4 Computational Cost

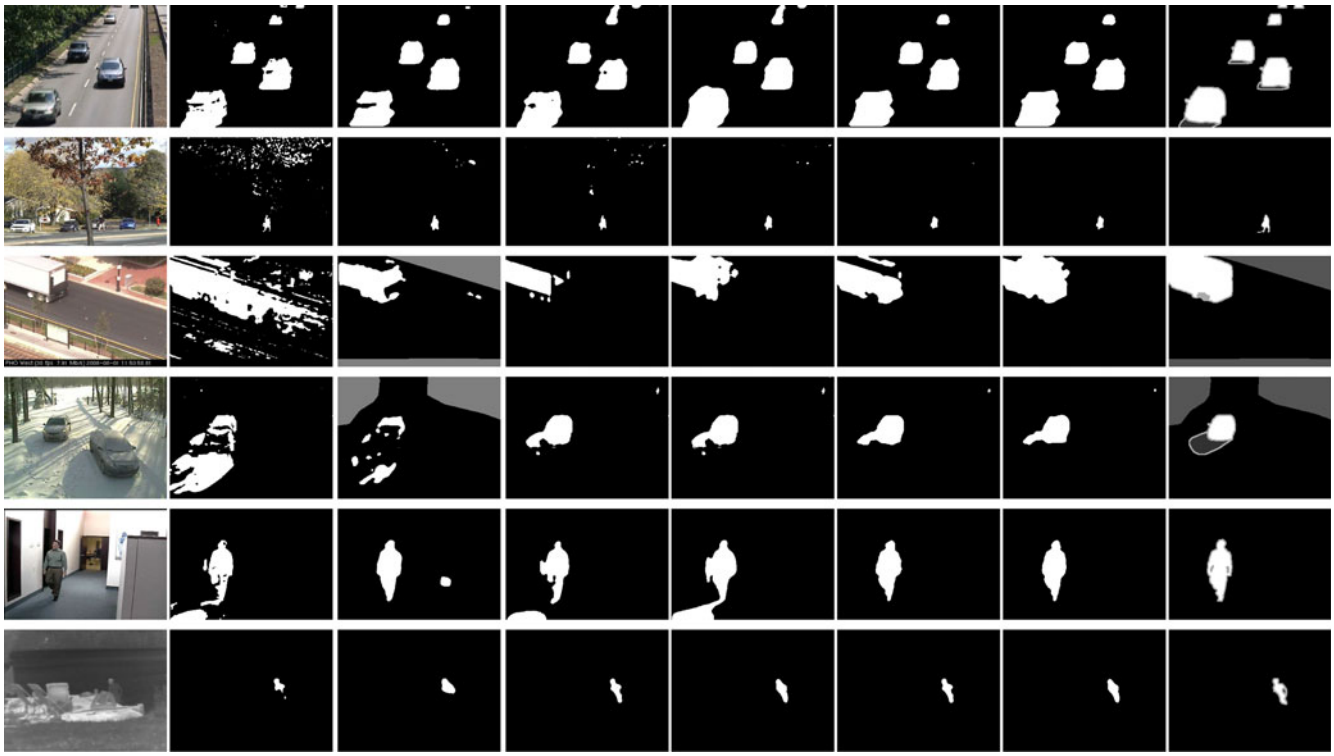
The proposed algorithms are evaluated on a machine with a 2.3 GHz Intel Core i7 CPU and 4 GB memory. Similar to [58], the runtime of the proposed algorithms is evaluated with respect to the background model based on the GMM [8] in Table 3. The main additional computational cost of the proposed SSHBM is the use of the MST-based hierarchical GMM and M-smoother. The computational complexity of this whole superpixel-based hierarchical GMM and M-smoother is relatively low as discussed in Sections 3 and 2.2. The computational cost of the proposed spatiotemporally-consistent background subtraction algorithm is much higher due to the use of optical flow. Nevertheless, real-time performance (more than 15 frames per second) can be achieved for QVGA videos with a Tesla K40 GPU.

5 Conclusion

Background subtraction is a fundamental research problem in computer vision. While pixel-based background models process each pixel independently and efficiently, these methods are not robust to noise due to sudden illumination changes. Although region-based background models can better describe scene changes, such approaches are less robust to frequent appearance variations. We propose effective and efficient background subtraction models based on hierarchical superpixel segmentation and robust estimator to exploit the strength of the two approaches. The proposed algorithms are robust to both frequent and sudden changes of pixel values as demonstrated by the performance evaluation on both SABS and ChangeDetection datasets against the state-of-the-art methods.

Acknowledgments

This work was supported in part by a GRF grant from the Research Grants Council of Hong Kong (RGC Reference: CityU 122212), a research grant from City University of Hong Kong (Project



(a) Input Frame (b) GMM [8] (c) PAWCS [63] (d) SBM (e) STBM (f) SSHBM (g) STSHBM (h) GT

Fig. 3. Sample background subtraction results. From top to bottom: video frames extracted from the *baseline*, *dynamic background*, *camera jitter*, *intermittent object motion*, *shadow* and *thermal* categories. (a) sample video frames. (b) to (g) background subtraction results obtained from the GMM, state-of-the-art PAWCS, proposed SBM, STBM, SSHBM and STSHBM. (h) ground-Truth (GT) foreground masks are shown. The proposed algorithms perform favorably against the other methods on the ChangeDetection dataset.

No. 9360153), the NSF CAREER Grant #1149783 and NSF IIS Grant #1152576.

REFERENCES

[1] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vis. Image Understanding*, vol. 122, pp. 4–21, 2014.

[2] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Comput. Sci. Rev.*, vol. 11/12, pp. 31–66, 2014.

[3] T. Bouwmans, A. Sobral, S. Javed, S. Jung, and E. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Comput. Sci. Rev.*, vol. 23, pp. 1–71, 2017.

[4] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Trans. Intell. Technol.*, vol. 1, pp. 43–60, 2016.

[5] A. Y. D. Zamalieva and J. Davis, "A multi-transformational model for background subtraction with moving cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 803–817.

[6] M. Narayana, A. Hanson, and E. Learned-Miller, "Coherent motion segmentation in moving camera videos using optical flow orientations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1577–1584.

[7] A. M. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.

[8] C. Stauffer and E. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1999, pp. 2246–2252.

[9] A. M. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using non-parametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[10] B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation: Applications to background modeling," in *Proc. Asian Conf. Comput. Vis.*, 2004, pp. 818–823.

[11] Z. Zivkovic and F. Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, pp. 773–780, 2006.

[12] D. Comaniciu, Y. Zhu, and L. Davis, "Sequential kernel density approximation and its application to real-time visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1186–1197, Jul. 2008.

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

[14] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2012, pp. 38–43.

[15] T. Parag, A. M. Elgammal, and A. Mittal, "A framework for feature selection for background subtraction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1916–1923.

[16] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.

[17] S. Zhang, H. Yao, and S. Liu, "Dynamic background modeling and subtraction using spatio-temporal local binary patterns," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1556–1559.

[18] S. Yoshinaga, A. Shimada, H. Nagahara, and R. Taniguchi, "Object detection using local difference patterns," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 216–227.

[19] B. Han and L. Davis, *Adaptive Background Modeling and Subtraction: A Density-Based Approach with Multiple Features*. Boca Raton, FL, USA: CRC Press, 2010.

[20] C. Silva, T. Bouwmans, and C. Frélicot, "An eXtended center-symmetric local binary pattern for background modeling and subtraction in videos," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 395–402.

[21] Y. Lin, Y. Tong, Y. Cao, Y. Zhou, and S. Wang, "Visual-attention based background modeling for detecting infrequently moving objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1208–1221, Jun. 2017.

[22] S. Varadarajan, P. Miller, and H. Zhou, "Region-based mixture of Gaussians modelling for foreground detection in dynamic scenes," *Pattern Recognit.*, vol. 48, no. 11, pp. 488–3503, 2015.

[23] P. Jodoin, V. Saligrama, and J. Konrad, "Behavior subtraction," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4244–4255, Sep. 2012.

[24] N. Jacobs and R. Pless, "Shape background modeling: The shape of things that came," in *Proc. IEEE Workshop Motion Video Comput.*, 2007, pp. 27–27.

[25] H. Bhaskar, L. Mihaylova, and A. Achim, "Video foreground detection based on symmetric alpha-stable mixture models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 8, pp. 1133–1138, Aug. 2010.

[26] H. Bhaskar, L. Mihaylova, and S. Maskell, "Background modeling using adaptive cluster density estimation for automatic human detection," in *Proc. 3rd German Workshop Sensor Data Fusion: Trends Solutions Appl.*, 2007, pp. 130–134.

[27] J. Lim and B. Han, "Generalized background subtraction using superpixels with label integrated motion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 173–187.

[28] D. Giordano, F. Murabito, S. Palazzo, and C. Spampinato, "Superpixelbased video object segmentation using perceptual organization and location prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4814–4822.

[29] S. Javed, S. H. Oh, A. Sobral, T. Bouwmans, and S. K. Jung, "Background subtraction via superpixel-based online matrix decomposition with structured foreground constraints," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 930–938.

[30] C. Zhao, T. Zhang, Q. Huang, X. Zhang, D. Yang, and S. Huang, "Background subtraction based on superpixels under multi-scale in complex scenes," in *Proc. Chinese Conf. Pattern Recognit.*, 2016, pp. 392–403.

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

[32] T. Tanaka, A. Shimada, R. Taniguchi, T. Yamashita, and D. Arita, "Towards robust object detection: Integrated background modeling based on spatio-temporal features," in *Proc. Asian Conf. Comput. Vis.*, 2009, pp. 201–212.

[33] B. Han and L. S. Davis, "Density-based multifeature background subtraction with support vector machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1017–1023, May 2012.

[34] T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of Gaussians for foreground detection-a survey," *Recent Patents Comput. Sci.*, vol. 1, no. 3, pp. 219–237, 2008.

[35] D. Pokrajac and L. Latecki, "Spatiotemporal blocks-based moving objects identification and tracking," *IEEE Visual Surveillance Performance Eval. Tracking Surveillance (VS-PETS)*, pp. 70–77, Oct. 2003.

[36] X. Fang, W. Xiong, B. Hu, and L. Wang, "A moving object detection algorithm based on color information," *J. Physics: Conf. Series*, vol. 48, pp. 384–387, 2006.

[37] D. Zhou and H. Zhang, "Accurate segmentation of moving objects in image sequence based on spatio-temporal information," in *Proc. Int. Conf. Mechatronics Autom.*, 2006, pp. 543–548.

[38] K. Schindler and H. Wang, "Smooth foreground-background segmentation for video processing," in *Proc. Asian Conf. Comput. Vis.*, 2006, pp. 581–590.

[39] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, "A fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Proc. Int. Conf. Multimedia Signal Process.*, 2012, pp. 177–185.

[40] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1402–1409.

[41] L. Bao, Q. Yang, and H. Jin, "Fast edge-preserving patchmatch for large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3534–3541.

[42] M. Chen, Q. Yang, Q. Li, G. Wang, and M. Yang, "Spatiotemporal background subtraction using minimum spanning tree and optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 521–534.

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

[44] 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. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 1–8.

[45] M. Pollack, "The maximum capacity through a network," *Operations Res.*, vol. 8, pp. 733–736, 1960.

[46] T. Hu, "The maximum capacity route problem," *Operations Res.*, vol. 9, pp. 898–900, 1961.

[47] L. Bao, Y. Song, Q. Yang, H. Yuan, and G. Wang, "Tree filtering: Efficient structure-preserving smoothing with a minimum spanning tree," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 555–569, Feb. 2014.

[48] C. Chu, I. Glad, F. Godtliebsen, and J. Marron, "Edge-preserving smoothers for image processing," *J. Amer. Statist. Assoc.*, vol. 93, no. 442, pp. 526–541, 1998.

[49] X. Wei, Q. Yang, Y. Gong, M. Yang, and N. Ahuja, "Superpixel hierarchy," *arXiv preprint arXiv:1605.06325*, 2016.

[50] M. Mares, "Two linear time algorithms for MST on minor closed graph classes," *Archivum Mathematicum*, vol. 40, pp. 315–320, 2004.

[51] A. Shimada, D. Arita, and R. Taniguchi, "Dynamic control of adaptive mixture-of-Gaussians background model," in *Proc. Int. Conf. Video Signal Based Surveillance*, 2006, Art. no. 5.

[52] N. McFarlane and C. Schofield, "Segmentation and tracking of piglets in images," *Mach. Vis. Appl.*, vol. 8, no. 3, pp. 187–193, 1995.

[53] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[54] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Understanding*, vol. 80, no. 1, pp. 42–56, 2000.

[55] L. Li, W. Huang, I. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 2–10.

[56] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imag.*, vol. 11, no. 3, pp. 172–185, 2005.

[57] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[58] A. Shimada, H. Nagahara, and R. Taniguchi, "Background modeling based on bidirectional analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1979–1986.
[59] M. Sedky, M. Moniri, and C. Chibelushi, "Object segmentation using full-spectrum matching of albedo derived from colour images," US Patent 2 374 109 12.10, 2011.
[60] M. D. Gregorio and M. Giordano, "A WiSARD-based approach to CDnet," in *Proc. BRICS Congr. Comput. Intell. 11th Brazilian Congr. Comput. Intell.*, 2013, pp. 172–177.
[61] N. Wang, T. Yao, J. Wang, and D. Yeung, "A probabilistic approach to robust matrix factorization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 126–139.
[62] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.
[63] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 990–997.