Spatiotemporal Background Subtraction Using Minimum Spanning Tree and Optical Flow*

Mingliang Chen[1], Qingxiong Yang[1,**], Qing Li[1], Gang Wang[2], and Ming-Hsuan Yang[3]

[1] Department of Computer Science
Multimedia software Engineering Research Centre (MERC)
City University of Hong Kong, Hong Kong, China
MERC-Shenzhen, Guangdong, China
[2] Nanyang Technological University, Singapore
[3] University of California, Merced

Abstract. Background modeling and subtraction is a fundamental research topic in computer vision. Pixel-level background model uses a Gaussian mixture model (GMM) or kernel density estimation to represent the distribution of each pixel value. Each pixel will be process independently and thus is very efficient. However, it is not robust to noise due to sudden illumination changes. Region-based background model uses local texture information around a pixel to suppress the noise but is vulnerable to periodic changes of pixel values and is relatively slow. A straightforward combination of the two cannot maintain the advantages of the two. This paper proposes a real-time integration based on robust estimator. Recent efficient minimum spanning tree based aggregation technique is used to enable robust estimators like M-smoother to run in real time and effectively suppress the noisy background estimates obtained from Gaussian mixture models. The refined background estimates are then used to update the Gaussian mixture models at each pixel location. Additionally, optical flow estimation can be used to track the foreground pixels and integrated with a temporal M-smoother to ensure temporally-consistent background subtraction. The experimental results are evaluated on both synthetic and real-world benchmarks, showing that our algorithm is the top performer.

Keywords: Background Modeling, Video Segmentation, Tracking, Optical Flow.

1 Introduction

Background modeling is one of the most extensively researched topics in computer vision. It is normally used as a fundamental pre-processing step in many

* This work was supported in part by a GRF grant from the Research Grants Council of Hong Kong (RGC Reference: CityU 122212), the NSF CAREER Grant #1149783 and NSF IIS Grant #1152576.

** Corresponding author.

D. Fleet et al. (Eds.): ECCV 2014, Part VII, LNCS 8695, pp. 521–534, 2014.

vision tasks, including video-surveillance, teleconferencing, video editing, human-computer interface, *etc.* It has recently experienced somewhat of a new era, as a result of publically available benchmarks for performance evaluation. These benchmarks simplify the comparison of a new algorithm against all the state-of-the-art algorithms. Both artificial pixel-level evaluation data sets [5] and real-world region-level data sets obtained by human experts [11] are available for comprehensive evaluation.

This paper focuses on traditional background subtraction problem with the assumption of a static video camera. There are significant publications on this topic, and can be classified into three broad categories: pixel-level background subtraction [9], [27], [8] [12], [33], [7], [10], region-level background subtraction [23], [15], [31], [14], frame-level background subtraction [32] and hybrid background subtraction [29] [28], [26], [13].

It is well understood that pixel-based models like mixture of Gaussians fail in sudden illumination changes. Region-based models on the other hand are more robust to these changes and tend to be vulnerable to periodic changes of pixel values. This paper proposes a way to synergistically combine the two to create a state-of-the-art background subtraction system.

The Gaussian mixture background model [27] is adopted in this paper. It is used to obtain an initial background estimate at each individual pixel location. Efficient minimum spanning tree (**MST**) based aggregation technique [30] is then integrated with a robust estimator - M-smoother to refine the initial estimates for a spatially-consistent background subtraction solution. The refined background estimates are then used to update the Gaussian mixture models to model stochastic changes in the value of each pixel. The updated Gaussian mixture models are thus robust to both periodic and sudden changes of pixel values. Note that comparing with the original Gaussian mixture background model [27], the extra computational cost is the **MST** based M-smoother, which is indeed extremely efficient. It takes about 6 ms to process a QVGA (320×240) color image on a single core CPU. Optical flow estimation is further employed to extend the proposed **MST** based M-smoother to the temporal domain to enhance temporal consistency. Although optical flow estimation is traditionally believed to be slow, recent fast nearest neighbor field [3] based optical flow algorithms like EPPM [2] enables the whole background subtraction pipeline to run in near realtime on state-of-the-art GPU.

The paper is organized as follows: Section 2 gives a brief overview of the Gaussian mixture background model adopted in the paper and the details of the proposed background modeling and subtraction algorithms. Section 3 reports results supporting the claims that the algorithm is currently the strongest available on standard benchmarks [5], [11]. Section 4 concludes.

2 Background Subtraction

A brief overview of Gaussian mixture background model is given in Sec. 2.1 and the proposed Spatially-consistent and temporally-consistent Background Models are presented in Sec. 2.2 and 2.3, respectively.

2.1 Gaussian Mixture Background Model

Stauffer and Grimson [27] propose to model the values of an image pixel as a mixture of Gaussians for background estimation. A pixel is considered to be background only when at least one of the Gaussians of the mixture 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 can be represented as follows

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

where η is a Gaussian probability density function

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

μ_k^t and Σ_k^t are the mean value and the covariance matrix of the k-th Gaussian in the mixture at time t, respectively.

Each pixel has a total of K different Gaussian distributions. To adapt to illumination changes, the new pixel values from the following frames will be used to update the mixture model, as long as they can be represented by a major component of the model.

To handle background changes, Shimada *et al.* [26] propose to leverage information from a future period with an acceptable delay as 33 milliseconds (the duration of just one video frame). The use of the information observed in future image frames was demonstrated to improve the accuracy by about 30%.

2.2 Spatially-Consistent Background Modeling Based on Minimum Spanning Tree

The bidirectional GMM [26] has been demonstrated to be a very effective background model while being very efficient. However, each pixel is processed independently and thus is less robust to noise. Region-based background model uses local texture information around a pixel to suppress the noise but is relatively slow and vulnerable to periodic changes of pixel values.

This section assumes that connected pixels with similar pixel values shall have similar background estimates, and thus spatially-consistent background segmentation can be obtained. The similarity between every two pixels is then defined using the minimax path [24] between the two pixels by treating the video frame as a connected, undirected graph $G = (V, E)$. The vertices V are all the image pixels and the edges E are all the edges between the nearest neighboring pixels. Minimax path can identify region boundaries without high contrast and will not cross the boundary of thin-structured homogeneous object; and thus can preserve details. Additionally, minimax path can be efficiently extracted with the use of a minimum spanning tree (**MST**) [16]. Recent study show that a minimum spanning tree can be extracted from an 8-bit depth image in time linear in the number of image pixel [1].

Let $d(p,q)$ denote the minimax path between a pair of node $\{p,q\}$ for the current frame I^t and $b_p^t = \{0,1\}$ denote the corresponding binary background estimates at pixel p obtained from Gaussian mixture background model. Minimax path $d(p,q)$ $(= d(q,p))$ is then employed in a robust estimator - M-smoother [6] to handle outliers in the coarse estimates from mixture of Gaussians. The refined background estimates is

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

When $\alpha = 1$, Eq. (3) is indeed a weighted median filter that utilize the minimax path length and thus is aware of the underlying regularity of the video frame. Because $b_p^t = \{0,1\}$, $(b_p^t)^\alpha = b_p^t$ and

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

Let $\mathcal{B}^t$ denote an image whose pixel value is (b_q^t) at pixel q and $\mathcal{F}^t$ denote an image whose pixel value is $|1 - b_q^t|$ at pixel q for frame t. Let

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

and

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

denote the weighted aggregation result of image $\mathcal{B}^t$ and $\mathcal{F}^t$, respectively. Eq. (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 new background estimate $b_p^{t,spatial}$ obtained from the proposed MST-based M-smoother will be used with the original estimate b_p^t to adjust the K Gaussian distributions, and the only difference is that the distributions will remain unchanged if either $b_p^{t,spatial}$ or b_p^t classifies pixel p as a foreground pixel. The noisy contribution from background pixel values for updating distributions can be significantly reduced using the spatially-consistent background estimates. As shown in Fig. 1 (b), part of the moving vehicle on the bottom right will be continuously detected as the background using Gaussian Mixture Model by adding foreground colors as new Gaussian distributions. Proposed MST-based M-smoother uses b_p^{new} as a new constrain to update Gaussian distributions and thus can correct most of the errors as can be seen in Fig. 1 (c).

(a)Video frame (b)Mixture of Gaussians[27] (c)Spatially-consistent

Fig. 1. Spatially-consistent background subtraction. (a) is a video frame extracted from the SABS data set [5] and (b) and (c) are foreground masks obtained from Gaussian mixture background model and the proposed spatially-consistent background model, respectively.

A Linear Time Solution. According to Eq. (7), the main computational complexity of the proposed proposed M-smoother resides in the weighted aggregation step in Eq. (5) and (6). The brute-force implementation of the nonlocal aggregation step is very slow. Nevertheless, the recursive matching cost aggregation solution proposed in [30] can be adopted:

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

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

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

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

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

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

The computational complexity is now straightforward. Only a total of two addition operations and three multiplication operations are required at each pixel location; and thus is extremely efficient.

2.3 Temporally-Consistent Background Modeling Based on Optical Flow

This section extends the spatially-weighted M-smoother proposed in Sec. 2.2 to the temporal domain as follows:

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

where the similarity measurement

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

$W(p, q_j)$ is obtained directly from optical flow estimation with the assumption that the background estimate for the same object appearing in difference video frames should be identical. Theoretically, the most robust optical flow should be employed to obtain the best performance. However, most of the optical flow algorithms are slow. According to Middlebury benchmark statistics, an optical flow algorithm takes around 1 minute to process a VGA resolution video frame. As a result, to ensure practicality, EPPM [2] which is currently fastest optical flow algorithm is used in this paper. Although it is not the top performer on standard benchmarks, EPPM significantly improves the accuracy of the proposed background subtraction algorithm as discussed in Section 3.

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

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

Eq. (12) can be simplified as follows:

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

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

The direct implementation of Eq. (15) is extremely slow as optical flow estimation will be required between any two video frames, that is optical flow estimation are required for a total of $\frac{t(t-1)}{2}$ image pairs 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 Eq. (14) so that optical flow estimation is required only between every two successive frames:

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

Spatiotemporal Background Modeling. A spatiotemporal background modeling solution can be directly obtained from Eq. (15) by replacing b_p^t with the spatially-consistent background estimates $b_p^{t,spatial}$ (from Section 2.2) in Eq. (17):

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

3 Experimental Results

In this section, the effectiveness of the proposed background subtraction method is experimentally verified for a variety of scenes using two standard benchmarks that use both artificial pixel-level evaluation data set [5] and real-world region-level data set obtained by human experts [11]. Visual or quantitative comparisons with the traditional model and recent methods are presented.

3.1 Evaluation Data Sets

Two public benchmarks containing both artificial and real-world scenes with different types of challenges were used for performance evaluation.

The first benchmark is SABS (Stuttgart Artificial Background Subtraction) [5], which is used for pixel-level evaluation of background models. Six artificial data sets used this benchmark cover a wide range of detection challenges. The *Dynamic Background* data set contains periodic or irregular movement in background such as waving trees or traffic lights; the *Bootstrapping* data set has no initialization data, thus subtraction starts after the first frame; the gradual scene change by varying the illumination constantly requires the segmentation when the contrast between background and foreground decreases in the *Darkening* data set; suddenly change are simulated in the *Light Switch* data set; the *Noisy Night* data set is severely affected by sensor noise which need to be coped with. Each data set contains 600 frames with the exception of *Darkening* and *Bootstrapping* both having 1400 frames. The sequences have a resolution of 800×600 pixels and are captured from a fixed viewpoint.

The second benchmark is ChangeDetection [11], which provide a realistic, camera-captured (no CGI), diverse set of videos. The real data sets used in this benchmark are representative of typical indoor and outdoor visual data captured today in surveillance, smart environment, and video database scenarios. A total of 31 video sequences with human labeled ground truth are used for testing. Similar to SABS benchmark, the video sequences are separated into six categories based on different types of challenges. The *Baseline* category represents a mixture of mild challenges typical of the other categories; The *Dynamic Background* category depicts outdoor scenes with strong (parasitic) background motion; The *Camera Jitter* category contains videos captured by unstable (e.g., vibrating) cameras; The *Shadows* category contains videos exhibiting strong as well as faint shadows; The *Intermittent Object Motion* category contains videos with scenarios known for causing "ghosting" artifacts in the detected motion,

i.e., objects move, then stop for a short while, after which they start moving again; The *Thermal* category contains videos captured by far-infrared cameras that result in typical thermal artifacts.

3.2 Evaluation Metric

The performance of an algorithm is evaluated on pixel-level, and the segmentation result of each pixel is a binary classification. The evaluation metric considers TP, FP and FN factors, where TP and FP denotes correctly and incorrectly classified foreground pixels respectively, FN denotes foreground pixels in GT are incorrectly classified background pixels. It also uses the F1-measure, a balance measure between precision and recall rate:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \tag{19}$$

$$F_1 = 2\frac{Recall \cdot Precision}{Recall + Precision}. \tag{20}$$

The F1-Measures (averaged over sequence) and Precision-recall charts of the performance of the approach with varying threshold will be computed and compared with respect to different data sets.

3.3 Evaluation on SABS Benchmark

This section reports performance evaluation of proposed background subtraction algorithm on SABS benchmark [5]. The maximal F-measure of the proposed spatially-consistent background model and the extended spatiotemporal background model are presented and compared with nine other background models reported to the benchmark in Table 1.

The proposed models clearly outperform all the other models on this benchmark. The proposed spatially-consistent background model outperforms the latest bidirectional Case-based background model [26] in almost every data set as can be seen in table 1. Additionally, the extended spatiotemporal background model outperforms all the other models under all types of challenges.

The corresponding recall precision curves with respect to different challenges are presented in Fig. 2. As can be seen, the proposed spatiotemporal consistent background model obtains the highest recall ratio under the same precision level. The proposed spatiotemporal background model clearly outperforms the others under three challenges: dynamic background (*Dynamic Background* data set), sudden illumination changes (*Light Switch* data set) and sensor noise (*Noisy Night* data set). Note that region-based background models are fragile to the first challenge while pixel-level background models are not robust to last two; the proposed models synergistically combine the two to cope with all the challenges.

Table 1. F-measures for the *SABS* benchmark [5]. The best and the 2^{th} best performers are shown in red color and blue color, respectively. The last column presents the average F-measures. Note that the proposed spatially-consistent background subtraction algorithm outperforms the others on average, and the extended spatiotemporal algorithm outperforms all the other on all the six data sets with different types of challenges.

Approach	*Basic*	*Dynamic Background*	*Bootstrap*	*Darkening*	*Light Switch*	*Noisy Night*	**Average**
McFarlane[20]	0.614	0.482	0.541	0.496	0.211	0.203	0.425
Stauffer[27]	0.800	0.704	0.642	0.404	0.217	0.194	0.494
Oliver[22]	0.635	0.552	-	0.300	0.198	0.213	0.380
McKenna[21]	0.522	0.415	0.301	0.484	0.306	0.098	0.354
Li[18]	0.766	0.641	0.678	0.704	0.316	0.047	0.525
Kim[17]	0.582	0.341	0.318	0.342	-	-	0.396
Zivkovic[33]	0.768	0.704	0.632	0.620	0.300	0.321	0.558
Maddalena[19]	0.766	0.715	0.495	0.663	0.213	0.263	0.519
Barnich[4]	0.761	0.711	0.685	0.678	0.268	0.271	0.562
AtsushiShimada[26]	0.723	0.623	0.708	0.577	0.335	0.475	0.574
Proposed (spatial)	0.764	0.747	0.669	0.672	0.364	0.519	0.623
Proposed (spatiotemporal)	**0.813**	**0.788**	**0.736**	**0.753**	**0.515**	**0.680**	**0.714**

3.4 Evaluation on ChangeDetection Benchmark

This section evaluates the proposed method using a real-world region-level benchmark - *ChangeDetection* with data sets obtained by human experts. Due to the lack of pixel-level accuracy in the ground-truth labels, a post-processing step like median filter is normally required for all background subtraction algorithms. As a result, the MST-based M-smoother proposed in Sec. 2.2 were applied to our background subtraction results as a post-processing step.

Table 2 presents the detailed evaluation results of the proposed background subtraction models on different types of challenges in terms of F-measure. Note that the proposed methods outperform the state of the art on this benchmark, especially when the *Shadow* category is excluded. The performance of the proposed models is good for most of the categories, especially on *Baseline*, *Camera Jitter*, *Intermittent Object Motion* and *Thermal* categories. Both of the proposed models are the either the best or second best performer on these four categories. Some of the extracted foreground mask are presented in Fig. 3 for visual evaluation.

The performance of the proposed method is surprisingly low on the *Shadow* category as visible in Table 2 and Fig. 4. This is because as we believe that *shadow deserves to be processed separately* and thus is not considered in the proposed background models. The performance on *Shadow* category can be greatly improved with the use of an existing shadow detection algorithm like [25].

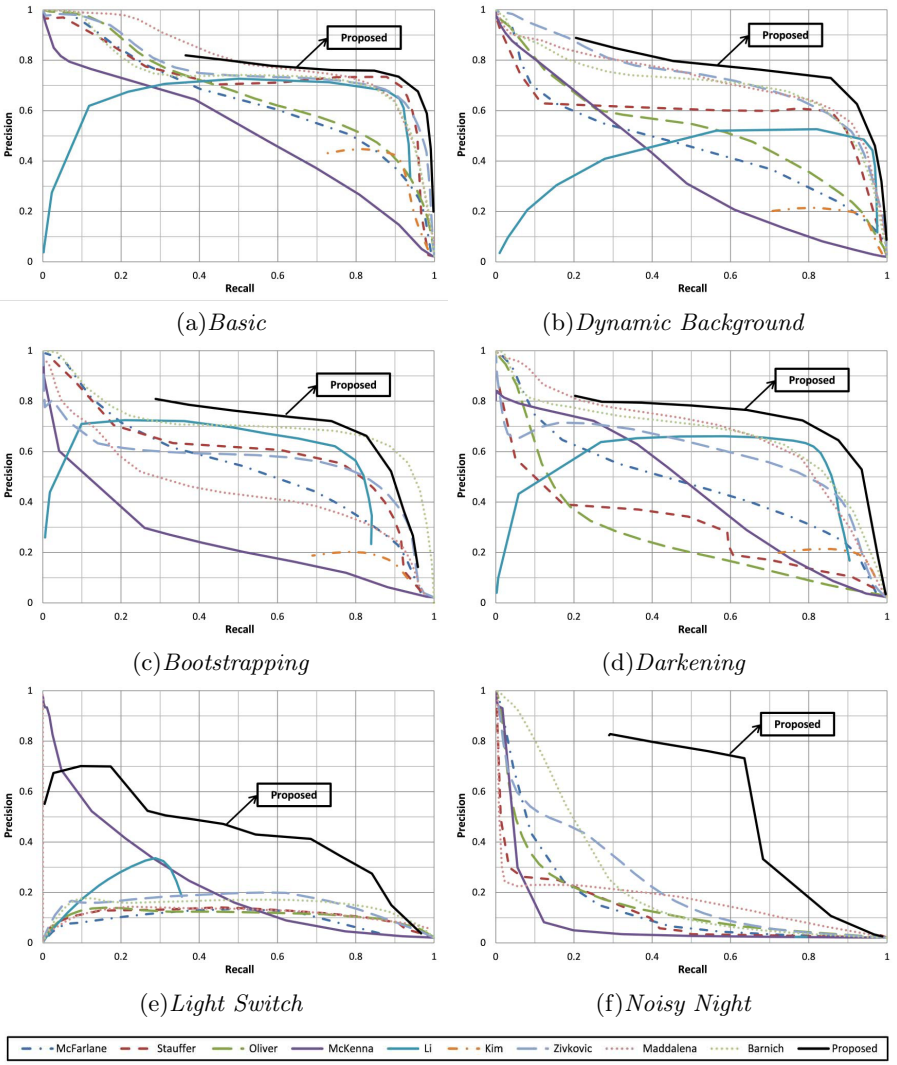


Fig. 2. Precision-recall charts for SABS benchmark [5] with different challenges. The dark solid curve presents the performance of the proposed spatiotemporal background subtraction algorithm. Note that it outperforms all the others overall.

Table 2. F-measures for *ChangeDetection* benchmark. The best and the 2^{th} best performers are shown in red color and blue color, respectively. The last two columns present the average F-measures including and excluding the *Shadow* category. Note that the proposed spatially-consistent background subtraction algorithm is comparable to the state-of-the-art algorithms, and the extended spatiotemporal algorithm outperforms all the others on average when shadows detection is required. However, the improvement is not significant. This is mainly because shadow modeling is not included in the proposed algorithms as we believe that shadow detection deserves to be considered separately. The improvement over the current state of the art is more significant when the *Shadow* category is excluded as shown in the last column.

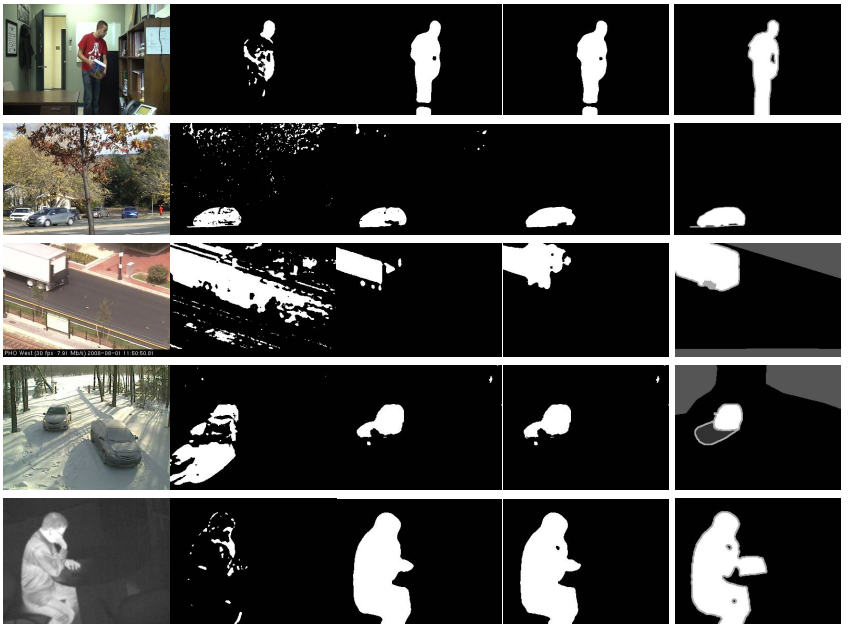
Approach	*Baseline*	*Dynamic Background*	*Camera Jitter*	*Intermittent Object Motion*	*Shadow*	*Thermal*	**Average** (*Shadow*)	**Average** (no *Shadow*)
PBAS-PID	0.9248	0.7357	0.7206	0.6267	0.8617	0.7622	0.7720	0.7540
DPGMM	0.9286	0.8137	0.7477	0.5418	0.8127	0.8134	0.7763	0.7690
Spectral-360	0.9330	0.7872	0.7156	0.5656	0.8843	0.7764	0.7770	0.7556
CwisarD	0.9075	0.8086	0.7814	0.5674	0.8412	0.7619	0.7780	0.7654
GPRMF	0.9280	0.7726	**0.8596**	0.4870	**0.8889**	0.8305	0.7944	0.7755
Proposed (spatial)	0.9250	0.7882	0.7413	0.6755	0.7606	0.8423	0.7888	0.7945
Proposed (spatiotemporal)	**0.9345**	**0.8193**	0.7522	**0.6780**	0.7764	**0.8571**	**0.8029**	**0.8082**

3.5 Computational Cost

This section reports the computational cost of the proposed background modeling and subtraction algorithms in Table 3. The proposed approach are tested on a laptop computer with a 2.3 GHz Intel Core i7 CPU and 4 GB memory. Similar to [26], the runtime of the proposed algorithms were evaluated with respect to GMM [27]. Comparing to the bidirectional GMM [26], the main additional cost of the proposed spatially-consistent background model is the use of the proposed MST-based M-smoother. Luckily, the computational complexity of this M-smoother is extremely low as has been analysis in Sec. 2.2. The total computational cost is higher than [26] but has a higher performance. The computational cost of proposed spatiotemporally-consistent background model is much higher due to the use of optical flow which is known to be slow. Nevertheless, near real-time performance (over 12 frames per second) can be obtained for QVGA-resolution videos when a state-of-the-art GPU is available.

Table 3. Computational cost of the proposed background modeling algorithms for QVGA-resolution videos (milliseconds/frame)

Method	GMM [27]	Bidirectional GMM [26]	Proposed (spatial)	Proposed (spatiotemporal)	
	CPU			CPU	GPU
Time	12	5	15	982	83



(a)Input frame (b)GMM[26] (c)Proposed (spatial) (d)Proposed (spatiotemporal) (e)Ground truth

Fig. 3. Visual comparison using foreground mask. From top to bottom: video frames extracted *Baseline*, *Dynamic Background*, *Camera Jitter*, *Intermittent Object Motion* and *Thermal* categories. (a) are the video frames extracted from different categories, (b) to (d) are the corresponding foreground masks obtained from GMM, proposed spatially-consistent and spatiotemporally-consistent background models, respectively and (e) are the ground-truth masks. As can be seen, the proposed extensions obviously outperforms the original GMM algorithm.

(a)Input frame (b)GMM [26] (c)Proposed (spatial) (d)Proposed (spatiotemporal) (e)Ground truth

Fig. 4. Visual comparison on *Shadow* category using foreground mask. GMM and the proposed background models do not detect shadows and thus cannot separate shadow from foreground.

4 Conclusions

In this paper, a background modeling and subtraction algorithm based on **MST** and optical flow estimation was proposed. The **MST** is used to form an efficient weighted M-smoother to enhance the spatial consistency while optical flow estimation is used to track the motion of image pixels to extend the proposed **MST** based M-smoother to the temporal domain.

Our algorithm is outperforming all other algorithms on both *SABS* and *ChangeDetection* benchmarks, but there is space left for improvement. For instance, our algorithm simply adopts the currently fastest optical flow algorithm [2] to ensure that the proposed algorithm is practical. However, other optical flow algorithms that are relatively slow but more accurate have not yet tested. They can potentially increase the performance of the proposed spatiotemporal background subtraction algorithm. Another question that was left for further study is how to adjust the algorithm for a moving camera.

References

1. Bao, L., Song, Y., Yang, Q., Yuan, H., Wang, G.: Tree filtering: Efficient structure-preserving smoothing with a minimum spanning tree. IEEE Transactions on Image Processing (2014)
2. Bao, L., Yang, Q., Jin, H.: Fast edge-preserving patchmatch for large displacement optical flow. In: CVPR (2014)
3. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.: Patchmatch: a randomized correspondence algorithm for structural image editing. TOG (2009)
4. Barnich, O., Droogenbroeck, M.V.: Vibe: A powerful random technique to estimate the background in video sequences. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (2009)
5. Brutzer, S., Hoferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: CVPR (2011)
6. Chu, C., Glad, I., Godtliebsen, F., Marron, J.: Edgepreserving smoothers for image processing. Journal of the American Statistical Association (1998)
7. Comaniciu, D., Zhu, Y., Davis, L.: Sequential kernel density approximation and its application to real-time visual tracking. PAMI 30(7), 1186–1197 (2008)
8. Elgammal, A.M., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using non-parametric kernel density estimation for visual surveillance. Proceedings of the IEEE (2002)
9. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
10. Elqursh, A., Elgammal, A.: Online moving camera background subtraction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 228–241. Springer, Heidelberg (2012)
11. Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P.: changedetection.net: A new change detection benchmark dataset. In: IEEE Workshop on Change Detection at CVPR (2012)
12. Han, B., Comaniciu, D., Davis, L.: Sequential kernel density approximation through mode propagation: Applications to background modeling. In: ACCV (2004)

13. Han, B., Davis, L.S.: Density-based multifeature background subtraction with support vector machine. PAMI 34(5), 1017–1023 (2012)
14. Han, B., Davis, L.: Adaptive Background Modeling and Subtraction: A Density-Based Approach with Multiple Features. CRC Press (2010)
15. Heikkila, M., Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. PAMI 28(4), 657–662 (2006)
16. Hu, T.: The maximum capacity route problem. Operations Research (1961)
17. Kim, K., Chalidabhongse, T., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging 11(3) (2005)
18. Li, L., Huang, W., Gu, I., Tian, Q.: Foreground object detection from videos containing complex background. In: Int. Conf. on Multimedia, pp. 2–10 (2003)
19. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. IEEE Transactions on Image Processing 17(7) (2008)
20. McFarlane, N., Schofield, C.: Segmentation and tracking of piglets in images. Machine Vision and Applications 8(3), 187–193 (1995)
21. McKenna, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Computer Vision and Image Understanding 80(1) (2000)
22. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. PAMI 22(8), 831 (2000)
23. Parag, T., Elgammal, A.M., Mittal, A.: A framework for feature selection for background subtraction. In: CVPR (2006)
24. Pollack, M.: The maximum capacity through a network. Operations Research (1960)
25. Prati, A., Mikic, I., Trivedi, M., Cucchiara, R.: Detecting moving shadows: Formulation, algorithms and evaluation. PAMI 25(7), 918–924 (2003)
26. Shimada, A., Nagahara, H., Taniguchi, R.: Background modeling based on bidirectional analysis. In: CVPR, pp. 1979–1986 (2013)
27. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking. In: CVPR, pp. 2246–2252 (1999)
28. Tanaka, T., Shimada, A., Taniguchi, R.-I., Yamashita, T., Arita, D.: Towards robust object detection: Integrated background modeling based on spatio-temporal features. In: Zha, H., Taniguchi, R.-I., Maybank, S. (eds.) ACCV 2009, Part I. LNCS, vol. 5994, pp. 201–212. Springer, Heidelberg (2010)
29. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principle and practice of background maintenance. In: ICCV (1999)
30. Yang, Q.: A non-local cost aggregation method for stereo matching. In: CVPR, pp. 1402–1409 (2012)
31. Yoshinaga, S., Shimada, A., Nagahara, H., Taniguchi, R.-I.: Object detection using local difference patterns. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 216–227. Springer, Heidelberg (2011)
32. Zhang, S., Yao, H., Liu, S.: Dynamic background modeling and subtraction using spatio-temporal local binary patterns. In: ICIP (2008)
33. Zivkovic, Z., Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27, 773–780 (2006)