Proceedings of the 2014 IEEE
International Conference on Robotics and Biomimetics
December 5-10, 2014, Bali, Indonesia

A Background Subtraction Algorithm for a Pan-Tilt Camera

Ying Chen and Hong Zhang

Abstract— **This paper is concerned with the detection of moving objects using a pan-tilt camera and a background subtraction algorithm. Traditionally, motion compensation is performed on the current image to align its pixels with their background models in previous frames. Pixel misalignment however can occur during motion compensation. Although this problem can be alleviated by using pixel motion such as the optic flow, motion information itself can be inaccurate and, together with pixel misalignment, contributes to false positive foreground detection. In this paper, we exploit the fact that pixel misalignment and inaccurate optic flow tend not to occur simultaneously for a pixel. Consequently, we can substantially improve the performance of the background subtraction algorithm by evaluating the marginal statistical models of appearance and motion separately – rather than jointly – in classifying whether a pixel is foreground. We will use experiments to validate our approach and establish its superiority to other competing algorithms in the literature.**

I. INTRODUCTION

Detection of moving objects from a static scene represents an important area of computer vision research. A common approach is to perform background subtraction, which identifies the moving objects by ”subtracting” the observed image from the estimated background. Areas with large differences are referred to as the foreground, i.e., the moving objects. Background can be modelled parametrically, e.g., Gaussian mixture model (GMM) [13], or non-parametrically, e.g., Support Vector Machines (SVMs) [7]. In a recent interesting study, ViBe [1] proposes to learn the background model non-parametrically by estimating the appearance distribution of a pixel from a sample of its spatial neighbours with a single frame, thus considerably speeding up the process of adaptation. This is a significant development for background subtraction with a moving camera as it allows immediate, undelayed background model adaptation to handle an image region that just enters the camera view.

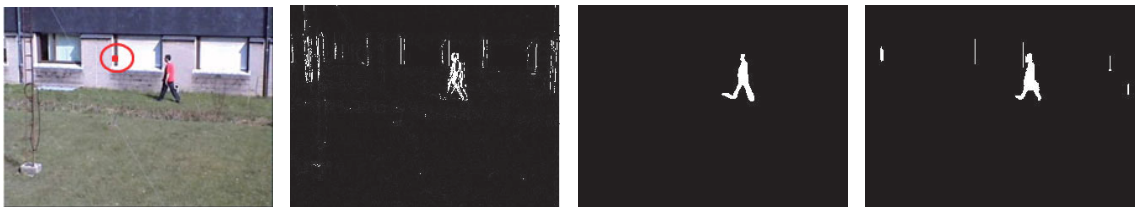
Traditional background subtraction approaches assume that the camera is static, which severely limits their usage in the moving platforms, such as cell phones, PTZ cameras, etc. Efforts have been made to extend background subtraction to moving cameras by compensating the camera motion via image registration. The consecutive frames, including the background model image, are registered to the same coordinate system, thus conventional background subtraction approaches can be applied to detect the moving objects. Images can be registered either globally [6], [8] or locally [2]. In global approaches, a homography is computed for compensating camera motion, i.e., aligning the background model and the current image before applying the conventional background subtraction. In local approaches, an expectation-maximization (EM) framework is employed to iteratively switch between motion estimation and background subtraction to detect moving objects. Precise image registration is not trivial and pixels can be misaligned. False alarms usually occur on the misaligned pixels which are compared to the unmatched background models, deteriorating the performance of background subtraction (see the green line in Fig. 2), e.g., a misaligned pixel (marked as red surrounded by a red circle) in Fig. 1 (a), is falsely labeled as foreground due to the large difference to its unmatched background model.

This problem can be alleviated by incorporating motion information of the foreground and background pixels [12], [3], [8], [5], [15], (see the blue and red lines in Fig. 2), the latter assumed to be stationary. While these approaches can reduce false positives generated from imperfect image registration, additional false positives or false negatives may result due to inaccurate optical flow estimation. For example, the motion of a pixel in the grass area (surrounded by a red circle) is incorrectly large due to the lack of texture leading to a false detection. These inaccuracies can result in either false positives (when motion estimate is incorrectly large) or false negatives (when motion estimate is incorrectly small).

In this paper, we present a background subtraction algorithm that is intended to suppress significantly false positive detections caused by pixel misalignment during motion compensation and by inaccurate motion estimation when motion is used in foreground classification, while introducing minimum false negatives (failures to detect foreground pixels). Our algorithm exploits the key observation that pixel misalignment and inaccurate motion estimation tend to occur not simultaneously or concurrently at a pixel. Consequently, we propose to classify a pixel by using its appearance and motion models separately rather than using the joint model of appearance and motion as in the literature [5], [9]. We can minimize the detrimental effect of spurious false negatives caused by this conservative classification rule – as experienced in our previous work [15] – by imposing spatial constraints on pixel labels in an MRF framework, to be solved via the graph-cut algorithm. We show that our proposed algorithm considerably outperforms the competing algorithms in terms of overcoming false positive detections and, at the same time, is comparable to those algorithms in terms of false negatives.

The rest of this paper is organized as follows. Related work is given and our background subtraction approach is discussed in detail in Section II. We compare the performance of our algorithm to existing state-of-art approaches in Section III. In Section IV, we conclude the paper and discuss possible extensions as future work.

978-1-4799-7397-2/14/$31.00 © 2014 IEEE

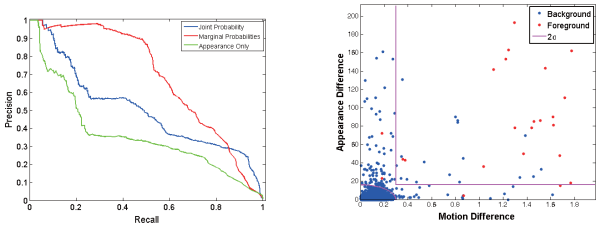


(a) False positives generated by pixel misalignment

(b) False positives generated by inaccurate motion estimate

Fig. 1. Problems in existing background subtraction methods from a pan-tilt camera. (a) From left to right: the current frame; image difference between the warped image and the current frame; the ground truth; the detection result. (b) From left to right: the current frame; the motion magnitude of the current frame: the darker the intensity, the smaller the motion magnitude; the ground truth; the detection result.

(a)



(b) (c)

Fig. 2. A simple example: illustrate two types of outliers that background subtraction approaches may fail to deal with: a) the outliers from pixel misalignment produced by imprecise image registration; b) the outliers from inaccurate motion estimate if motion is incorporated.

II. Materials and Methods

A. State of the Art

Background Subtraction approaches that exploit motion and appearance cues can be divided into three categories: a) sparse-to-dense approaches where motion is used for sparse labelling [12], [3], b) approaches where motion and appearance are incorporated jointly [5], [9], and c) approaches where two cues are incorporated but used separately [15]. The intuition of sparse labelling is that movements of moving objects and the background objects obey different geometric constraints which can tell them apart. However, these approaches may discard objects with short-term trajectories and may fail to detect small objects due to insufficient data points.

Instead of making use of geometric constraints, pixel-wise motion can be computed and incorporated with appearance cue either jointly or separately, which can be explained with the popular adaptive Gaussian mixture model (GMM) [13]. Given the appearance I of a pixel at location (x, y) of the image, its probability of being background is described:

$$P(I_t(x,y)) = \sum_{i=1}^{K} w_{i,t} N(I_t \mid \mu_{i,t}, \Sigma_{i,t}) \tag{1}$$

where K is the number of Gaussians in the mixture weighted by $w_{i,t}$, and $\mu_{i,t}$ and $\Sigma_{i,t}$ are their means and covariance matrices. For simplicity of discussion and without loss of generality, we use a single Gaussian to describe the background model and drop the unnecessary subscripts and the reference to location (x, y), i.e.,

$$P(I) = N(I \mid \mu_I, \Sigma_I) \tag{2}$$

where the new subscript I refers to appearance (intensity or color) of the pixel. As mentioned, to improve the performance of background subtraction, motion feature can be added to the background model [5], i.e., one can use the joint probability – assuming independence between I and M –

$$P(I, M) = P(I)P(M) = N(I \mid \mu_I, \Sigma_I) N(M \mid \mu_M, \Sigma_M) \tag{3}$$

to compute the likelihood of a pixel being background where M refers to a motion feature.

For those approaches that jointly model motion and appearance, the classification of a pixel is based on thresholding $P(I, M)$ or evaluating the exponent of the Gaussian probability in Equation (3). That is, a pixel is a background if

$$(I - \mu_I)^T \Sigma_I^{-1} (I - \mu_I) + (M - \mu_M)^T \Sigma_M^{-1} (M - \mu_M) \leq Th \tag{4}$$

where Th is a threshold (typically 2.5 in [13]). As one can see from Equation (3), in case of an outlier pixel, i.e., a background pixel that exhibits a different appearance from its model due to misalignment or large motion due to incorrect optic flow estimation, a small $P(I)$ or small $P(M)$ would result, and $P(I, M)$ would be small, leading to a false positive foreground pixel. In other words, in case of misalignment, I would significantly differ from μ_I or in case of inaccurate motion estimation, M would from μ_M, so that the threshold test in Eq.(4) would fail, causing a false foreground pixel.

Instead of jointly modelling motion and appearance, [15] and [4] filter out the false positives, i.e., the falsely-labeled unmoved background pixels, generated from appearance based approaches using the motion cue, since only those foreground, i.e., moving pixels, share uniform motion [4] or relatively large motion magnitude [15]. The intuition is that the outlier background pixels exhibit the property that one of the two features (appearance or motion) is correct whereas the other wrong (see the blue pixels along the vertical and horizontal axes, which is proved by the experimental data (see Fig. 2)). This could be due to the fact that misalignment is caused by one algorithm (e.g., image warping via a global homograhpy [6]), whereas motion inaccuracy is caused by a separate algorithm ([14] in our case). In fact, these approaches consider a classification procedure in which marginal probabilities of the background model are applied individually to label a pixel, and in which a background label according to either feature is sufficient to overule the other, independently of the label that it produces. Such a classification procedure corresponds to the rectangular decision boundary in Fig. 2 (c), which is able to overcome many false positive foreground pixels committed by a decision boundary defined by Eq.(4) and approximated by the elliptical curve near the origin. Specifically, the labelling procedure of marginally incorporating motion and appearance cues can be generalized as:

if $(I-\mu_I)^T\Sigma_I^{-1}(I-\mu_I) \leq Th_I$ or
$(M-\mu_M)^T\Sigma_M^{-1}(M-\mu_M) \leq Th_M$ **then**
$pixel \leftarrow background$
else
$pixel \leftarrow foreground$
end if

However, as these approaches consider pixels individually, they may be too conservative and thus lower the number of retrieved foreground pixels, i.e., the recall rate. For example, a pixel with inaccurate motion estimate could become false negative when the motion estimate is incorrectly small though it may look different from the background appearance model. In fact, neighbouring pixels should share similar motion or appearance as long as they are not on the boarders (of the moving objects), and thus should be given the same label. As will be seen, we will reduce these false negatives effectively through the use of spatial Markov constraints in the Markov random field (MRF) framework implemented through graphcut.

We should point out the idea of using the marginal distributions was explored in our previous work [15], with two key differences. First, in our previous work we dealt with a stationary camera and used the classic GMM background model, whereas in the current work we use the ViBe framework [1] to achieve undelayed background model initialization, in order to handle a moving camera that introduces new background region in each frame. Second, our pixel classification algorithm is embedded in the MRF framework in order to impose spatial constraints between pixel labels. As a result, our present algorithm is able to reduce false negatives and outperforms our previous work, as will be seen in the next section.

B. Graph Model

Marginally incorporating motion and appearance cues can be implemented as two one-layer graph models without considering topological constraints, one of which models appearance, the other models motion, and thus the detection result can be achieved by anding the graphs' labelling solved by graph cuts. In fact, labelling from the two graphs should satisfy some topological constraints. Intuitively, a pixel should achieve identical labels from the perspective of both appearance and motion. Moreover, neighbouring pixels should also share the same label if they are similar to each other in terms of either motion or appearance cues. Without considering these constraints while anding the labels generated from two features may end up with false negatives which get correct labels from one of the layers. Instead of building two one-layer graph models, we exploit the marginal probabilities of appearance and motion via a two-layer graph model where two types of spatial constraints are imposed to reduce false positives at the minimum expense of false negatives.

Here, we assume a pinhole camera model where a pan-tilt camera is mounted at a fix-point where the depth of object in the scene are not changed, and thus reduce the parallax-translation problem. To implement our algorithm in the MRF framework, we create a graph of two layers, one based on appearance (L_a) and the second based on motion (L_m). Weights on the edges from a pixel to the source and sink nodes are defined according to its background model ($P(I)$ or $P(M)$), whereas the weights between neighbouring nodes (in a layer or between layers) are given in terms of the imposed spatial constraints accordingly. Then we solve for labels of the pixels using the standard s-t graphcut algorithm. The detection result can be achieved by anding the labellings from the two layers. In our graph model, we consider two types of constraints: label consistency constraint and spatial coherence constraint. Our graph model is shown in Fig. 3.

C. Spatial Constraints

Label consistency constraint implies that the labels of two corresponding nodes in two layers should be identical. This constraint is satisfied for foreground pixels, i.e., moving pixels are assumed to have different appearance and motion to background, and vice visa for background pixels. One

advantage of label consistency constraint is that we can lower the chance of miss-detections. Imposing label consistency constraint can push a cut to group the two corresponding nodes to the foreground cluster if the other cue has a strong confidence to be foreground, in return, maintaining a high recall. This constraint can be imposed by connecting corresponding nodes of the two layers with a constant weight, which prevent a cut through these links so that the corresponding nodes can maintain the same label.

Spatial coherence constraint derives from the pairwise Markov property, i.e., a variable is conditionally independent of all variables given its neighbours. In graph cut, it can be interpreted as any two neighbouring pixels with high similarity should keep the same label while the discontinuity should be preserved. This constraint can be imposed via links between neighbouring pixels in a layer given weights based on their similarity.

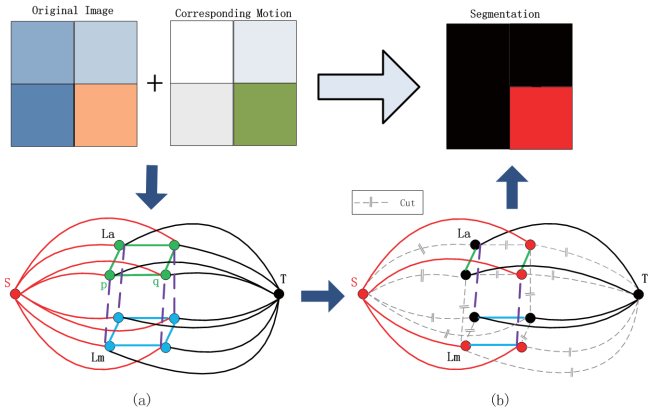


Fig. 3. Our graph model

D. Background Modelling

In the appearance layer, in order to achieve undelayed adaptation, we implement $P(I)$ non-parametrically with the framework introduced in ViBe [1]. Since $P(M)$ does not require adaptation, $P(I, M)$ can be updated in an undelayed manner, an important property for handling new pixels entering the view of a pan-tilt camera.

Normally, pixel p, with its appearance I_p (here we use the RGB colour space) can be classified as background if the probability of belonging to background is high. In ViBe, this probability can be measured as the number of votes that I_p is close to its collection of background samples $\mathbf{B}_p$. Thus, the label of p depends on whether the number of its background samples that are within a distance R to I_p is larger than or equal to a given threshold.

Instead of classifying pixels by comparing the observation and the background model which might not be good at dealing with inaccuracies of image registration, we leave the classification to graph cut. In appearance layer, we compute the confidence of belonging to its background as the similarity of pixel p to its closest background sample. Mathematically, $Pr(I_p|BG)$ is defined as:

$$Pr(I_p|BG) = \max_{s_j \in \mathbf{B}_p} exp(-\frac{(I_p - s_j)^2}{2R^2}) \quad (5)$$

In the motion layer, a background pixel should be static after compensating the camera motion, while a foreground pixel, i.e., a pixel belonging to a moving object, has a non-zero motion. Considering the noise of motion estimation, the background motion model can be formulated as Gaussian distribution $\mathcal{N}(\mathbf{m}_i; \mathbf{0}, \mathbf{\Sigma})$ with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma}$.

III. Experimental Results

In this section, we compare our approach to three state-of-the-art background subtraction (BGS) methods qualitatively and quantitatively, i.e., Appearance-based BGS (ABGS) [11], BGS that Marginally incorporates motion and appearance cues (MBGS) [15], and BGS that Jointly models these two cues (JBGS) [5]. The difference matrix between these methods is given in Table I. Quantitative evaluation is made at both the pixel level – using precision, recall, and F-measure – and at the object level using the metrics proposed in [10]: correct detections (CD), detection failures (DF) and false alarm (FA), which are measured with respect to foreground regions or connected components. The video sequences, (two captured outdoors, two indoors) which we use for testing, involve camera panning and tilting. One is from the recent work of ViBe [1] with 1137 frames and the others from our own experimental setup using a PTZ camera by Axis Communications Only camera motion in pan and tilt is allowed, while the camera focal length (zoom) is fixed. Zoom operation (including zoom-in and zoom-out) is not considered since the one-to-many or many-to-one pixels mapping in image registration due to the change of focal length is not our concern in this paper.

First of all, qualitative results are shown in Fig. 4. For each video sequence, we select two representative frames for showing the performance. Column (a) is the input video frame being processed, and Column (b) is the manually annotated ground truth. Columns (c) through (e) are the results from [11], [5], [15], and Column (f) is the result of our algorithm. Clearly, our approach is superior to all three competing methods in terms of false positive detections, and similar to them in terms of true positive detections.

The result of the quantitative evaluation is given in Table II, computed by averaging the performance metrics over all the images in each video sequence. First of all, at the pixel level, our algorithm enjoys high precision similar recall, and the significantly higher F-measure value, which combines precision and recall, than the competing methods. Second, at the region level, all competing algorithms have similar correct detection (CD) rates, a metric that is similar to recall at the pixel level. However, our algorithm leads to a significantly lower false alarm (FA) rate, a measure that is related to precision. Finally, our algorithm is highly competitively in terms of detection failure (DF). These results clearly demonstrate the superiority of our approach,

TABLE I

SUMMARIZATION OF OUR APPROACH AND THREE OTHER APPROACHES FOR COMPARISON. "MARGINALLY " AND "JOINTLY" REFER TO AS THE TWO WAYS OF INCORPORATING APPEARANCE CUE AND MOTION CUE IN BACKGROUND MODELLING, "A" AND "M" REFER TO APPEARANCE AND MOTION RESPECTIVELY, "SC" IS THE ABBREVIATION OF "SPATIAL CONSTRAINTS".

	A	M	Marginally	Jointly	SC
ABGS	Y				
MBGS	Y	Y	Y		
JBGS	Y	Y		Y	Y
Ours	Y	Y	Y		Y

TABLE II

QUANTITATIVE COMPARISON OF COMPETING METHODS USING FOUR VIDEO SEQUENCES IN FIG. 4 AT BOTH THE PIXEL LEVEL USING PRECISION (P), RECALL (R), AND F-MEASURE (FM), AND AT THE OBJECT LEVEL USING CORRECT DETECTION (CD), FALSE ALARMS (FA) , AND DETECTION FAILURE (DF).

Outdoor Sequence 1						
	P	R	FM	CD	FA	DF
ABGS	0.46	0.79	0.58	**0.86**	0.87	**0.05**
MBGS	**0.90**	0.42	0.58	0.59	**0.02**	0.29
JBGS	0.59	**0.84**	0.69	0.87	0.57	0.08
Ours	0.85	0.74	**0.79**	0.82	0.15	0.08
Outdoor Sequence 2						
	P	R	FM	CD	FA	DF
ABGS	0.33	**0.93**	0.49	**0.95**	0.97	**0.03**
MBGS	**0.91**	0.70	0.79	0.39	**0.03**	0.50
JBGS	0.66	**0.93**	0.77	0.92	0.64	0.05
Ours	0.87	0.88	**0.88**	0.88	0.27	0.07
Indoor Sequence 1						
	P	R	FM	CD	FA	DF
ABGS	0.55	0.72	0.63	0.75	0.86	0.15
MBGS	**0.94**	0.52	0.67	0.44	**0.08**	0.45
JBGS	0.33	**0.93**	0.49	**0.98**	0.93	**0.01**
Ours	0.90	0.76	**0.83**	0.86	0.18	0.07
Indoor Sequence 2						
	P	R	FM	CD	FA	DF
ABGS	0.58	0.61	0.60	0.81	0.83	0.12
MBGS	**0.91**	0.32	0.48	0.44	**0.10**	0.46
JBGS	0.40	**0.89**	0.55	**0.96**	0.90	**0.02**
Ours	0.83	0.66	**0.74**	0.78	0.22	0.10

particularly with respect to reducing false positive detections that are caused by pixel misalignment and inaccurate optic flow computation, at a minimum expense of recall.

IV. CONCLUSIONS AND FUTURE WORKS

This paper proposes a background subtraction algorithm for detecting moving objects with a pan-tilt camera. A critical concern when applying background subtraction algorithm in this scenario is false positive detection caused by pixel misalignment during motion compensation and by inaccuracy in optic flow when motion information is used in foreground classification. We take advantage of the observation that false positive detections due to pixel misalignment and optic flow inaccuracy do not tend to occur simultaneously at the same pixel locations, and we thus propose to use the marginal statistically models of appearance and motion of a pixel independently, rather than the joint distribution of both motion and appearance, in making the detection decision. Experimental results using video sequences establish the superiority of our approach in both qualitative and quantitative terms.

The future work includes the extension to a hand-held camera where image registration is more complex and hard to deal with due to the depth variation. Another concern is to exploit our approach in the scenario where the focal length is changed.

REFERENCES

[1] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, 2011.

[2] S.A. Berrabah, G. De Cubber, V. Enescu, and H. Sahli. Mrf-based foreground detection in image sequences from a moving camera. *IEEE International Conference on Image Processing*, pages 1125 – 1128, 2006.

[3] Ali Elqursh and Ahmed Elgammal. Online moving camera background subtraction. *Europe Conference on Computer Vision*, 7577:228–241, 2012.

[4] H. Fradi and J. Dugelay. Robust foreground segmentation using improved gaussian mixture model and optical flow. *IEEE International Conference on Systems*, pages 248–253, 2012.

[5] M. Gong and L. Cheng. Incorporating estimated motion in real-time background subtraction. *IEEE International Conference on Image Processing*, 3265 - 3268, 2011.

[6] Eric Hayman and Jan-Olof Eklundh. Statistical background subtraction for a mobile observer. *IEEE International Conference on Computer Vision*, 1:67–74, 2003.

[7] Horng-Horng Lin, Tyng-Luh Liu, and Jen-Hui Chuang. A probabilistic svm approach for background scene initialization. *International Conference on Image Processing*, 3:893–896, 2002.

[8] Anurag Mittal and Dan Huttenlocher. Scene modeling for wide area surveillance and image synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:160–167, 2000.

[9] Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. *Computer Vision and Pattern Recognition*, 2:302–309, 2004.

[10] Jacinto Nascimento and Jorge Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, pages 761–774, 2006.

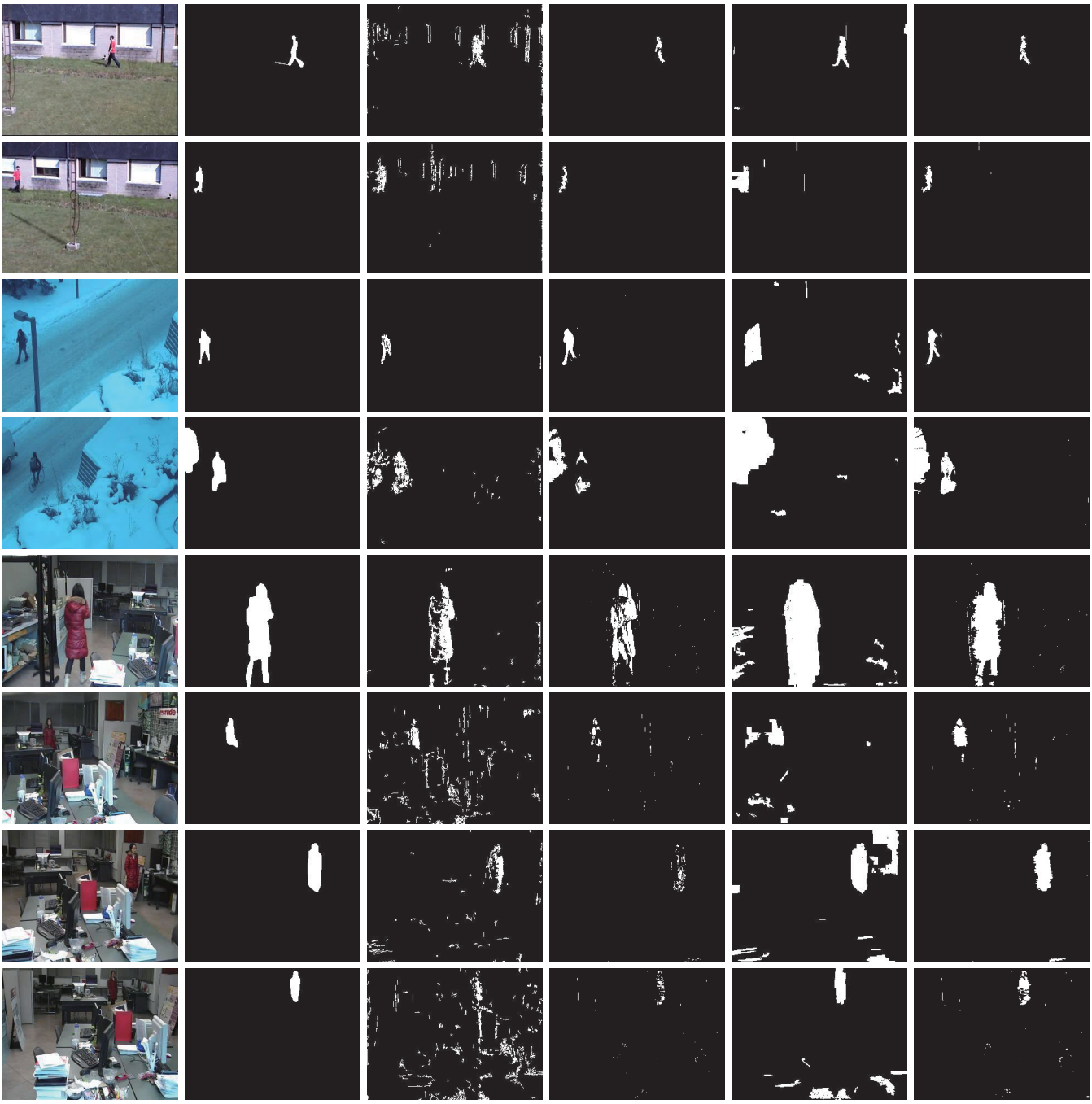
[11] Ying Ren, Chin-Seng Chui, and Yeong-Khing Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letteres*, 24:183–196, 2003.

[12] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving camera. *IEEE International Conference on Computer Vision*, pages 1219 – 1225, 2009.

[13] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[14] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l^1 optical flow. In *Proceedings of the 29th DAGM conference on Pattern Recognition*, pages 214–223, 2007.

[15] D. Zhou and H. Zhang. Modified gmm background modeling and optical flow for detection of moving objects. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, pages 2224–2229, 2005.



(a) Img (b) GT (c) ABGS [11] (d) MBGS [15] (e) JBGS [5] (f) Ours

Fig. 4. Comparison with three state-of-art background subtraction methods over four video sequences.