2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Stationary foreground detection for video-surveillance based on foreground and motion history images

Diego Ortego, Juan C. SanMiguel
Video Processing and Understanding Lab
Escuela Politécnica Superior, Universidad Autónoma de Madrid, SPAIN
Email: Diego.Ortego@estudiante.uam.es, Juancarlos.Sanmiguel@uam.es

Abstract

Stationary foreground detection is a common stage in many video-surveillance applications. In this paper, we propose an approach for stationary foreground detection in video based on the spatio-temporal variation of foreground and motion data. Foreground data are obtained by Background Subtraction to detect regions of interest. Motion data allows to filter out the moving regions and it is estimated using median filters over sliding windows. Spatio-temporal patterns of both data are computed through history images and the final detection is obtained using a two-threshold scheme that considers motion activity. Partial visibility of stationary foreground for short-time intervals is handled to increase robustness. The results over challenging video-surveillance sequences show an improvement of the proposed approach against the related work.

1. Introduction

Detecting stationary foreground regions in video has recently become an active area of research in many video-surveillance areas such as the detection of abandoned objects [1] and illegally parked vehicles [2]. This task remains unsolved for complex sequences such as crowded scenarios as it faces many challenges related with illumination changes, low resolution images, object occlusions, high density of moving objects (increasing the number of cast shadows) and initialization of the detection algorithms.

Common stationary foreground detectors are based on the background subtraction approach [3], which provides binary foreground maps. Some proposals focus on tracking foreground regions to detect the stationary ones [4][5]. They are limited as current tracking performance is only acceptable in situations with few moving objects [6]. Avoiding tracking, many pixel-wise approaches are proposed based on dual-backgrounds [6][7][8], accumulators [9][10], sub-sampling [11], specific object classifiers [12] or properties of background models [13][14]. However, background subtraction presents many false positives in crowds that decrease stationary detection performance. Recently, combinations between foreground and motion analysis have been investigated to address these limitations in crowds [15].

In this paper, we propose an approach for detecting stationary foreground regions in video that combines foreground and motion data. Building on the concept of History Images [16], we develop energy maps (images) that account for spatio-temporal patterns of foreground and motion. Foreground data are extracted using a standard background subtraction approach [17]. Motion data is obtained by computing frame differences in the nearby frames (before and after the analysis instant) using a median filter. Finally, both energy maps are combined through a two-threshold technique that considers the spatial location of motion activity. Occlusion handling is included at pixel level to tolerate partial visibility of stationary regions. The proposed approach is evaluated and compared on video-surveillance datasets in presence of detection challenges such as occlusions, illumination changes and clutter.

The structure of this paper is as follows. Section 2 discusses the related work. Section 3 describes the proposed approach. Experimental results are presented in Section 4. Finally, Section 5 summarizes the main conclusions.

2. Related work

Many approaches have been proposed for stationary region detection in video [3]. They can be classified into based on tracking [4][5] or background subtraction [10]. As tracking accuracy is significantly degraded in complex sequences, such as crowded videos, this section focuses on the second category that does not use tracking and can be applied to a wide variety of video-surveillance scenarios.

Two major error sources affect the performance of detection approaches based on background subtraction. The first corresponds to photometric factors (illumination changes, camouflages, shadows and reflections) whereas the second

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

derives from sequences with high density of moving objects (multiple occlusions and algorithm initialization). Adaptive background subtraction (ABS) has been proposed to handle photometric errors by continuously updating the background model [13]. Combinations of fast and slow adaptation rates can be used for stationary detection [6]. However, such adaptation might decrease detection performance as static objects can be incorporated into the background before they become static [12]. Thus, slow rates are preferred that reduce the robustness to photometric errors. Moreover, background initialization is complex in crowded sequences that, if incorrect, may lead to many false positives (of foreground), which decrease stationary detection performance.

In this context, several approaches have been developed based on temporal accumulation [9][10] and subsampling [11] of foreground masks. Moreover, modeling properties can be used such as the transitions of Gaussian Mixture Modeling (GMM) approach [13]. They can be extended by defining the states of foreground pixels through finite-state-machines such as for GMMs [14] and dual-backgrounds [8]. Recent results show that sampling approaches present best results with high spatial accuracy (less noise in the final mask) and low temporal accuracy (detection delay) [14]. However, selection of the sampling instants remains unsolved, which is critical for efficient analysis. All previous approaches are limited for crowded scenes as many false stationary detections are produced due to the high amount of detected foreground. Specific object classifiers can be used for solving this limitation in crowds [12]. However, it requires to know the objects of interest, which is often not available. Foreground sampling can be combined with inter-frame motion [15], demonstrating that motion could be used to remove false detections in crowds. However, it shares the drawbacks of sampling and the spatial accuracy dependency with camouflage errors.

In summary, no approach is able to perfectly perform in crowded scenes considering low false positive detection and spatio-temporal robustness of static mask. We propose an approach that combines the most relevant features of existing approaches avoiding the sub-sampling drawbacks.

3. Stationary region detection

In this section, we describe the proposed approach for stationary foreground detection. It comprises two analysis, both at pixel level on frame-by-frame basis, for foreground and motion data (see Figure 1). Each analysis has two stages to model spatio-temporal patterns: feature extraction and history image computation. The two resulting history images are combined to get an image representing the foreground-motion variation over time, which is thresholded to get the stationary foreground mask. Finally, occlusion handling is performed to recover lost pixels due to frequent object occlusions in crowds.

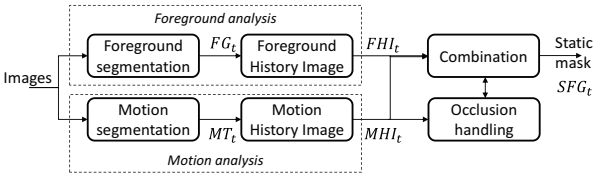

Figure 1. Overview of the proposed approach.

3.1. Foreground analysis

First, background subtraction is applied to detect foreground. We have used the method proposed in [17] due to its low computational cost and robustness to noise. Detection considers pixel neighborhood and it is summarized as:

$$FG_t(\mathbf{x}) \Longleftrightarrow \sum_{\mathbf{d}\in\mathcal{N}(\mathbf{x})} \left(I_t\left(\mathbf{d}\right) - B_t\left(\mathbf{d}\right)\right)^2 > \beta, \quad (1)$$

where $\mathbf{x}$ and $\mathbf{d}$ are pixel locations $\{x, y\}$; $\mathcal{N}(\mathbf{x})$ is an $N \times N$ patch centered at $\mathbf{x}$; I_t and B_t are current and background frames and β is a decision threshold. $FG_t(\mathbf{x}) = 1(0)$ indicates foreground (background) for the pixel located at $\mathbf{x}$.

Then, we measure the foreground temporal variation to get a Foreground History Image $FHI_t(\mathbf{x})$, considering foreground and background detections as follows:

$$FHI_t(\mathbf{x}) = FHI_{t-1}(\mathbf{x}) + w^f_{pos} \cdot FG_t(\mathbf{x}), \quad (2)$$

$$FHI_t(\mathbf{x}) = FHI_{t-1}(\mathbf{x}) - w^f_{neg} \cdot (\sim FG_t(\mathbf{x}))\,, \quad (3)$$

where $\sim$ is the logical NOT operation; w^f_{pos} and w^f_{neg} are two weights to manage the contribution of the foreground $(FG_t(\mathbf{x}) = 1)$ and background $(\sim FG_t(\mathbf{x}) = 1)$ detections. For giving a temporal sense to stationary detection, we should increase FHI_t values one-by-one ($w^f_{pos} = 1$) when they belong to foreground and reset FHI_t values to 0 when they are background. Nevertheless, temporally sparse errors in foreground detection may cause loosing correct stationary detections if reset to 0. This frequently happens in crowds where a static region is occluded by fast moving objects which cause camouflage errors. Hence, penalization weight w^f_{neg} should decrease FHI_t at a higher rate than positive one w^f_{pos} without resetting to 0 for increasing robustness against foreground errors (e.g., $w^f_{neg} = 15$).

Finally, the result of this analysis is $FHI_t(\mathbf{x})$, a foreground score that increases when the pixel is foreground and decreases when it belongs to the background model.

3.2. Motion analysis

Recent works show the use of motion information for filtering false positives caused by high densities of moving objects, thus helping to detect stationary regions [5][15]. Current use focuses on thresholding inter-frame differences and

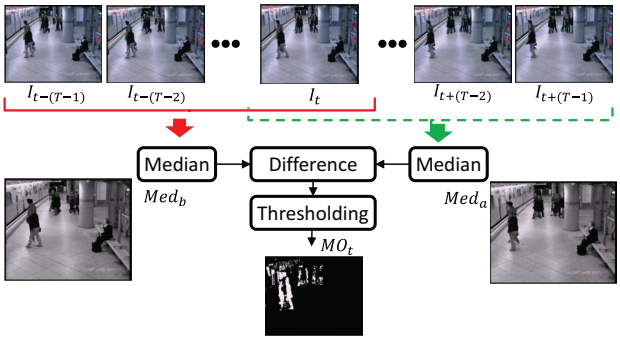

Figure 2. Motion extraction scheme using median filtering over temporal windows before and after the frame under analysis.

then, applying the sub-sampling approach over the temporal sequence of differences [15]. However, stationary regions are frequently occluded in presence of many moving objects and, therefore, the no-motion state (or static) is difficult to observe for such regions during all frames of a determined time interval. Hence, successful performance requires selecting the correct number and frequency of the samples taken, which can not be guaranteed for all situations.

We propose to solve these limitations by extending the motion analysis over temporal windows of length T (see Figure 2). Although multiple occlusions affect stationary regions in crowds, they usually last for few frames and the most predominant region in short-time intervals corresponds to the stationary one. Moreover, History Images [16] could be employed instead of sampling approach to avoid deciding when samples have to be taken.

For extracting motion using temporal windows, we apply a median filter before and after the frame under analysis:

$$Med_b = Median\{I_{t-T+1}, ..., I_t\} \quad (4)$$
$$Med_a = Median\{I_t, ..., I_{t+T-1}\} \quad (5)$$

where Med_a and Med_b are the median images of temporal windows of length T taken after and before I_t (all images at gray level). A delay is introduced for each instant t to get the next $T-1$ frames. The choice of T depends on the speed of objects and duration of occlusions, requiring high values for slow occlusions. Empirical testing over real sequences obtained good performance for T values ranging from 10 to 20. Then, final motion image is obtained as:

$$MO_t(\mathbf{x}) = \begin{cases} 1 & if \quad |Med_b - Med_a| < \tau \\ 0 & otherwise \end{cases}, \quad (6)$$

where $MO_t(\mathbf{x}) = 1$ is for absence of motion and τ is a threshold to set the no-motion case. We automatically get τ by applying the Kapur method [18] on $|Med_b - Med_a|$.

Finally, temporal variation of the no-motion mask $MO_t(\mathbf{x})$ is computed via the Motion History Image $MHI_t(\mathbf{x})$, which is similar to the foreground case:

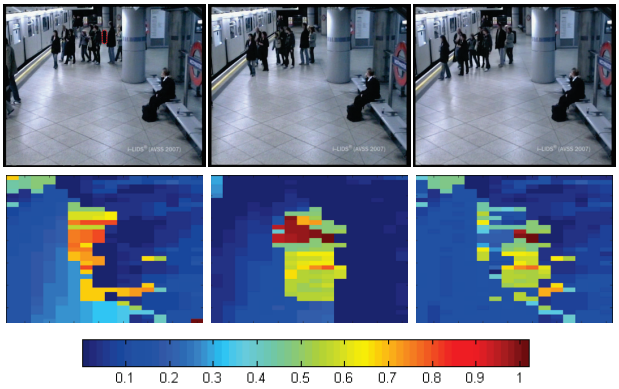

Figure 3. Example of $MHI_t(\mathbf{x})$ using the median-based proposed approach (PRO) and the standard frame difference (FD) [15]. First row: frames 875, 917 and 940 of sequence *AVSS07 Med*. Second row: $MHI_{917}(\mathbf{x})$ results of frame 917 using FD (left), PRO (center) and their absolute difference (right) for the red-dashed rectangle in frame 875 (suitcase). PRO estimates better the no-motion state of the stationary region with higher values in $MHI_{917}(\mathbf{x})$.

$$MHI_t(\mathbf{x}) = MHI_{t-1}(\mathbf{x}) + w^m_{pos} \cdot MO_t(\mathbf{x}), \quad (7)$$

$$MHI_t(\mathbf{x}) = MHI_{t-1}(\mathbf{x}) - w^m_{neg} \cdot (\sim MO_t(\mathbf{x})), \quad (8)$$

where w^m_{pos} and w^m_{neg} are two weights for controlling the contribution of the no-motion ($MO_t(\mathbf{x}) = 1$) and motion ($\sim MO_t(\mathbf{x}) = 1$) cases. Similarly to stationary detection using $FHI_t(\mathbf{x})$, we should increase $MHI_t(\mathbf{x})$ values one-by-one ($w^m_{pos} = 1$) when they belong to the no-motion state and reset $MHI_t(\mathbf{x})$ values to 0 when they belong to the motion state. We use this scheme as $MHI_t(\mathbf{x})$ is included to compensate high values of $FHI_t(\mathbf{x})$ caused by continuous motion of moving objects, only keeping $FHI_t(\mathbf{x})$ values of non-moving pixels. Hence, we set $w^m_{neg} = MHI_{t-1}(\mathbf{x})$ to reset $MHI_t(\mathbf{x})$ to 0 when motion is detected. Figure 3 depicts an example where the no-motion of the entire stationary region is only detected with the proposed $MHI_t(\mathbf{x})$.

Finally, the result of the motion analysis is $MHI_t(\mathbf{x})$, a no-motion score that increases when the pixel does not suffer motion or decreases when it suffers motion.

3.3. Combination

After obtaining $FHI_t(\mathbf{x})$ and $MHI_t(\mathbf{x})$, we normalize them to the range $[0, 1]$ considering the video framerate (fps) and the stationary detection time (t_{static}) :

$$\overline{FHI_t}(\mathbf{x}) = min\{1,\ FHI_t(\mathbf{x})/(fps \cdot t_{static})\}, \quad (9)$$
$$\overline{MHI_t}(\mathbf{x}) = min\{1,\ MHI_t(\mathbf{x})/(fps \cdot t_{static})\}. \quad (10)$$

Then, we compute the mean of both normalized images to get a stationary history image $SHI_t(\mathbf{x})$ representing foreground-motion variation over time. Finally, stationary detection mask is obtained by thresholding as:

Criteria	Non-crowded					Crowded													Total
	AVSS07	PETS06			PETS07	AVSS07		PETS07		PETS06						HALL			
	Easy	S7_C3	S4_C3	S4_C4	S5_C3	Med	Hard	S5_C1	S5_C2	S7_C1	S7_C4	S1_C1	S1_C4	S4_C1	S4_C2	H_S1	H_S2	H_S3	
Background Initialization	L	L	L	L	L	L	L	H	H	H	M	H	M	H	H	H	M	H	-
Illumination changes	L	-	-	-	M	L	L	M	M	-	-	-	-	-	-	L	L	L	-
Motion level	L	L	L	L	L	M	H	H	H	H	M	H	L	H	H	H	M	H	-
Overall complexity	L	L	L	L	L	H	H	H	H	H	M	H	M	H	H	H	M	H	-
Number of frames	4291	3401	3051	3051	2900	4834	5311	2900	2900	3401	3401	3021	3021	3051	3051	10000	10834	15102	87521
Annotated stationary regions	2	1	3	4	2	14	13	3	3	1	1	2	2	6	3	3	1	11	75

Table 1. Description of the sequences of the evaluation set. (Key. L:Low. M:Medium. H:High).

$$SFG_t(\mathbf{x}) =, \begin{cases} 1 & if \quad SHI_t(\mathbf{x}) \geq \eta \\ 0 & otherwise \end{cases}, \tag{11}$$

where $\eta \in (0, 1]$ is the threshold for stationary detection. Its value should be high ($\eta = 1$, if no foreground or motion errors). $FHI_t(\mathbf{x})$ and $MHI_t(\mathbf{x})$ must indicate the stationary regions to detect as they equally contribute to SHI_t being not possible such detection relying only on one of them.

However, stationary regions constantly occluded remain undetected in most of the situations as $\overline{MHI_t}(\mathbf{x})$ is not able to capture the required consecutive no-motion to allow the increase of its values. We include an additional condition that reduces η in pixels where $\overline{FHI_t}(\mathbf{x})$ has reached a high value with previous or current motion:

$$SFG_t(\mathbf{x}) =, \begin{cases} 1 & if \quad \overline{FHI_t}(\mathbf{x}) \geq \eta \,\&\, \overline{MHI_t}(\mathbf{x}) < \eta \,\& \\ & \quad SHI_t(\mathbf{x}) \geq \eta \cdot factorTh \\ 0 & otherwise \end{cases}, \tag{12}$$

where $factorTh \in (0, 1)$ weights the threshold η. It should have high (low) values for sequences presenting low (high) motion activity. Eq. 12 allows to apply a lower threshold to pixels with previous or current motion ($\overline{MHI_t}(\mathbf{x}) < \eta$) and high foreground history image values ($\overline{FHI_t}(\mathbf{x}) \geq \eta$), obtaining detections in situations with motion.In summary, a two-thresholding scheme is proposed that applies conditions to pixels with no-motion (Eq. 11) and motion in previous time instants (Eq. 12).

3.4. Occlusion handling

After detecting stationary regions, a reduction of $MHI_t(\mathbf{x})$ values might occur due to total or partial occlusions and, therefore, reducing the values of $SHI_t(\mathbf{x})$ to satisfy any of the conditions in Eqs. 11 and 12. We add an occlusion handling method to recover initial detections where they are lost. Unlike previous works [5][15], we focus on pixels instead at blob level for such handling as it more robust to foreground errors. For each pixel we check some conditions and propagate previous detections as follows:

$$SFG_t(\mathbf{x}) = \begin{cases} 1 & if \quad SFG_{t-1}(\mathbf{x}) = 1 \,\&\, \overline{MHI_t}(\mathbf{x}) < \eta \,\& \\ & \quad \overline{FHI_t}(\mathbf{x}) \geq \eta \cdot factorOc \\ 0 & otherwise \end{cases}, \tag{13}$$

where $factorOc \in (0, 1)$ is the tolerance to temporally sparse foreground errors (e.g., camouflages) that reduce $FHI_t(\mathbf{x})$ and cause loss of static detections. This recover is applied when pixels have experienced motion ($\overline{MHI_t}(\mathbf{x}) < \eta$) and high accumulated foreground ($\overline{FHI_t}(\mathbf{x}) \geq \eta \cdot factorOc$). The lower values of *factorOc*, the higher robustness against errors. However too low values delay the disappearance of static detections which no longer exist.

4. Experimental results

In this section, we present and compare the experimental results of the proposed approach.

4.1. Setup

Experiments are performed on selected sequences from AVSS2007[1], PETS2006[2] and PETS2007[3] datasets. They provide a diverse set of public video-surveillance scenarios (see Table 1). We also use a larger dataset for crowded situations recorded at a faculty hall (HALL). We manually annotated all stationary regions as ground truth.

To evaluate detection performance, we use standard Precision (P), Recall (R) and F-score (F) measures:

$$P = TP/(TP + FP), \tag{14}$$

$$R = TP/(TP + FN), \tag{15}$$

$$F = 2 \cdot P \cdot R/(P + R), \tag{16}$$

where TP, FP and FN are, respectively, correct, false and missed detections (as compared to ground-truth ones).

To set up the proposed approach, we use the common values for framerate (25 fps) and stationary detection time ($t_{static} = 20$ secs). We use the following values to guarantee that no static region appears before t_{static}: $w^f_{pos} = w^m_{pos} = 1$, $w^m_{neg} = MHI_{t-1}(\mathbf{x})$ and $\eta = 1$. Temporally sparse foreground errors (i.e., camouflages) are tolerated by empirically setting $w^f_{neg} = 15$ and $factorOc = 0.8$. Finally, after testing on crowded videos, we observed a decrease around 50-25% of $MHI_t(\mathbf{x})$ values with t_{static} =10-20 secs, thus, we set $factorTh = 0.625$ to

[1] http://www.avss2007.org/

[2] http://www.cvg.rdg.ac.uk/PETS2006/

[3] http://www.cvg.rdg.ac.uk/PETS2007/

Approach		Non-crowded						Crowded													
		AVSS07	PETS06			PETS07	Mean	AVSS07		PETS07		PETS06						HALL			Mean
		Easy	S7_C3	S4_C3	S4_C4	S5_C3		Med	Hard	S5_C1	S5_C2	S7_C1	S7_C4	S1_C1	S1_C4	S4_C1	S4_C2	H_S1	H_S2	H_S3	
[9]	P	.33	1	1	.80	.50	.72	.58	.48	.16	.20	.05	.12	.12	.33	.27	.05	.50	1	.34	0.32
	R	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.83	.33	1	1	1	0.93
	F	.50	**1**	**1**	**.88**	**.66**	.81	.73	.65	.28	.33	.10	.22	.22	.50	.41	.10	.67	**1**	.51	.44
[11]	P	.33	1	.75	.80	.50	.67	.51	.52	.17	.21	.05	.12	.12	.33	.27	.07	.30	.14	.37	0.24
	R	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.83	.33	1	**1**	1	0.93
	F	.50	**1**	.85	**.88**	**.66**	.78	.68	.68	**.29**	.35	.11	.22	.22	.50	.41	.10	.46	.25	.54	.37
[15]	P	.40	1	1	.80	.50	.74	.60	.58	0	.16	.14	.16	.10	.40	.38	.09	.60	.50	.55	0.33
	R	1	1	1	1	1	1	.85	.76	0	.66	1	1	.50	1	.83	.33	1	**1**	.91	0.53
	F	**.57**	**1**	**1**	**.88**	**.66**	**.82**	.70	.66	0	.26	**.25**	**.28**	.16	**.57**	.52	**.14**	**.75**	.67	.69	.43
Proposed	P	.33	1	1	.80	.50	.72	.66	.68	.17	.23	.1	.16	.13	.40	.41	.07	.60	1	.58	0.39
	R	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.83	.33	1	1	1	0.93
	F	.50	**1**	**1**	**.88**	**.66**	.81	**.80**	**.81**	**.29**	**.37**	.18	**.28**	**.23**	**.57**	**.55**	.10	**.75**	**1**	**.73**	**.51**

Table 2. Comparative results of the proposed approach using Precision (P), Recall (R) and F-score (F). Bold indicates best results.

consider the contributions of $FHI_t(\mathbf{x})$ and $MHI_t(\mathbf{x})$ (respectively, 100% and 25%) to $SHI_t(\mathbf{x})$ in crowds. The same parameters are used for all the experiments.

4.2. Results

Table 2 compares the proposed approach with the most popular ones based on foreground accumulation [9], subsampling [11] and foregound-motion sampling [15]. In non crowded sequences, results are very similar getting all high performance. The best results are obtained by [15] because there are not many occlusions, so it is able to avoid false detections through motion analysis without losing correct detections. For crowded sequences, many occlusions and high motion take place. Previous works [9][11] are not good enough in these situations, getting in general very high Recall values but low Precision ones (high false positive rate). [15] is capable to improve Precision in most of the sequences, because it eliminates many false detections. However, this filtering removes stationary detections in many cases, so Recall is also decreased counteracting the previous improvement. The proposed approach is able to maintain the stationary region detection rate (Recall) and still removing the false detections caused by high motion. Globally, the proposed approach has an improvement around 18% and 16% for, respectively, Precision and F-score as compared to the best results of the selected approaches.

Figure 4 shows some visual examples of the compared approaches. Examples of rows 1, 2, 4 and 5 show the performance improvement of the proposed approach removing false detections caused by high motion. Furthermore, unlike [15], examples 1, 2, 3 demonstrate how the proposed method is able to keep detections in the stationary mask, although a motion analysis is included for dealing with high density situations in [15]. All examples exhibit false detections caused by non-correct background models due to the high complexity for their initialization and other photometric factors (shadows and illuminations).

5. Conclusions

This paper has presented an approach for stationary foreground region detection. It computes spatio-temporal variations of foreground and motion data extracted from the video sequence. A two-threshold scheme is applied to combine the previous analysis and detect stationary regions. The results over heterogeneous datasets show that the proposed approach is effectively applied to crowded sequences outperforming related work and demonstrating the use of motion to remove false positive detections.

As future work, we will explore the use of complex models for foreground detection and background initialization, automatic tuning of algorithm parameters and the use of region-level information.

Acknowledgments

This work has been partially supported by the Spanish Government (TEC2011-25995 EventVideo).

References

[1] J. Ferryman, D. Hogg, J. Sochman, A. Behera, J. Rodriguez-Serrano, S. Worgan, L-Li, V. Leung, M. Evans, P. Cornic, S. Herbin, S. Schlenger, and M. Dose. Robust abandoned object detection integrating wide area visual surveillance and social context. *Pattern Recogn. Lett.*, in press, 2013. 1

[2] J.T. Lee, M. S. Ryoo, M. Riley, and J.K. Aggarwal. Real-time illegal parking detection in outdoor environments using 1-d transformation. *IEEE Trans. Circuits Syst. Video Technol.*, 19(7):1014–1024, July 2009. 1

[3] A. Bayona, J. C. SanMiguel, and J. M. Martínez. Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 25–30, Genova (Italy), Sep. 2009. 1, 2

[4] J.C. San Miguel and J.M. Martínez. Robust unattended and stolen object detection by fusing simple algorithms. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 18–25, Sept. 2008. 1, 2

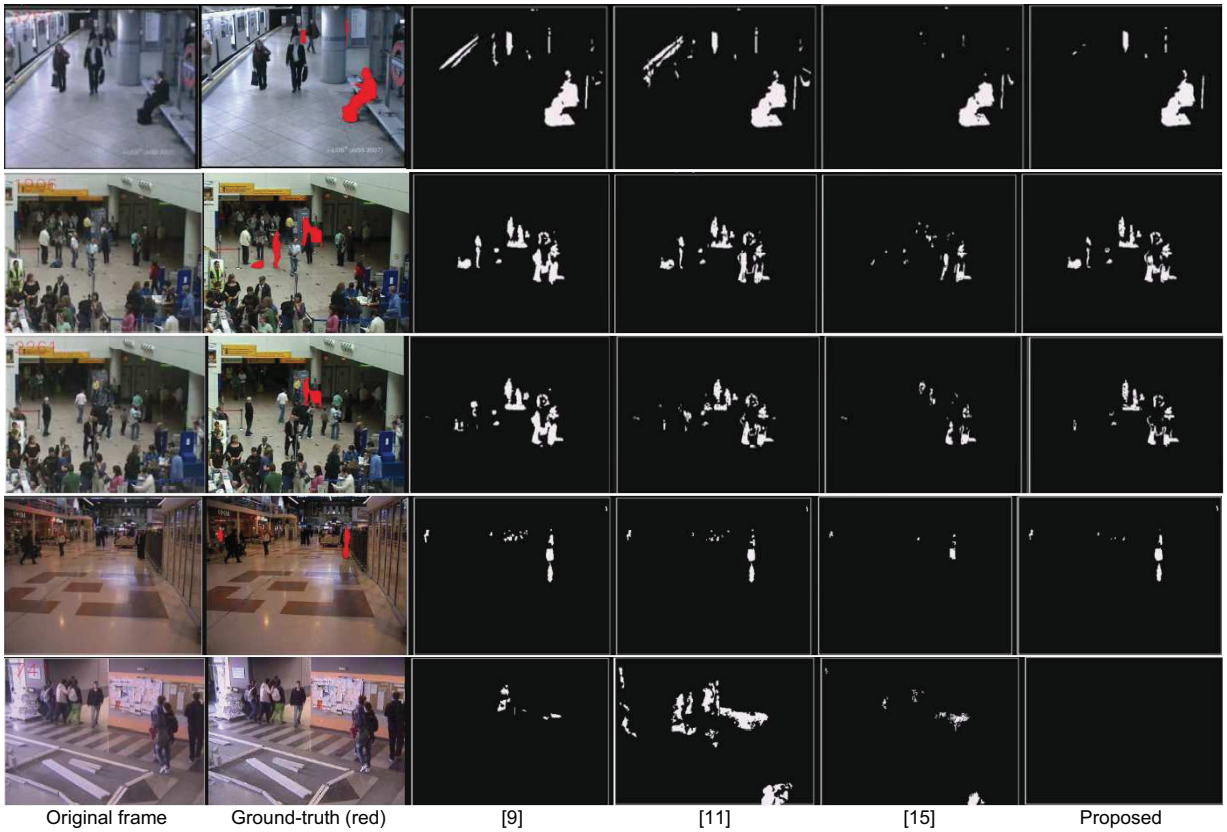

Figure 4. Sample results for stationary foreground detection for (from top to bottom row) *Hard*, *S5_C1*, *S5_C1*, *S4_C1* and *H_S2* sequences.

[5] J. Kim, B. Kang, H. Wang, and D. Kim. Abnormal object detection using feedforward model and sequential filters. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 70–75, Beijing (China), Sept. 2012. 1, 2, 3.2, 3.4

[6] F. Porikli, Y. Ivanov, and T. Haga. Robust abandoned object detection using dual foregrounds. *EURASIP J. Adv. Signal Process.*, Article ID 197875, 2008. 1, 2

[7] A. Singh, S. Sawan, M. Hanmandlu, V. K. Madasu, and B. C. Lovell. An abandoned object detection system based on dual background segmentation. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 352–357, Sep. 2009. 1

[8] R. Evangelio and T. Sikora. Static object detection based on a dual background model and a finite-state machine. *EURASIP J Image Video Process.*, Article ID 858502, 2011. 1, 2

[9] S. Guler and J. A. Silverstein. Stationary objects in multiple object tracking. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 248–253, Sept. 2007. 1, 2, 4, 4.2

[10] L. Maddalena and A. Petrosino. Stopped object detection by learning foreground model in videos. *IEEE Trans. Neural Netw. Learn. Sys*, (in press), 2013. 1, 2

[11] C. Jing-Ying, L. Huei-Hung, and C. Liang-Gee. Localized detection of abandoned luggage. *EURASIP J. Adv. Signal Process.*, Article ID 675784, 2010. 1, 2, 4, 4.2

[12] M. Bhargava, C. Chen, M.S. Ryoo, and J.K. Aggarwal. Detection of object abandonment using temporal logic. *Mach. Vision Appl.*, 20:271–281, 2009. 1, 2

[13] Y. Tian, A. Senior, and M. Lu. Robust and efficient foreground analysis in complex surveillance videos. *Mach. Vision Appl.*, 23(5):967–983, 2012. 1, 2

[14] Q. Fan and S. Pankanti. Modeling of temporarily static objects for robust abandoned object detection in urban surveillance. In *Proc. IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 36–41, Sep. 2011. 1, 2

[15] A. Bayona, J.C. SanMiguel, and J.M. Martínez. Stationary foreground detection using background subtraction and temporal difference in video surveillance. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, pages 4657–4660, Sept. 2010. 1, 2, 3.2, 3, 3.4, 4, 4.2

[16] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell*, 23(3):257–267, 2001. 1, 3.2

[17] A. Cavallaro, Steiger O., and Ebrahimi T. Semantic video analysis for adaptive content delivery and automatic description. *IEEE Trans. Circuits Syst. Video Technol.*, 15(10):1200–1209, Oct. 2005. 1, 3.1

[18] C. Su and A. Amer. A real-time adaptive thresholding for video change detection. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, pages 157–160, 2006. 3.2