MULTI-FEATURE STATIONARY FOREGROUND DETECTION FOR CROWDED VIDEO-SURVEILLANCE

Diego Ortego, Juan C. SanMiguel

Video Processing and Understanding Lab (VPULab), Universidad Autónoma de Madrid, Spain

ABSTRACT

We propose a novel approach for stationary foreground detection in crowds based on the spatio-temporal evolution of multiple features. A generic framework is presented to detect stationarity where history images model the spatio-temporal feature patterns. A feature is proposed based on structural information over each pixel neighborhood for dealing with shadows and illumination changes. A multi-feature detector is composed by combining the history images of three features (namely, foreground, motion and structural information) to estimate the foreground stationarity over time, which is later thresholded to detect stationary regions. Experimental results over challenging video-surveillance sequences show the improvement of the proposed approach against related work as structural information reduces false detections, which are common in crowded places.

Index Terms— Stationary foreground detection, structural similarity, illumination changes, shadows, video-surveillance

1. INTRODUCTION

Automatic video-surveillance is an active research area due to the increasing society's concern about security in public crowded areas. In this domain, detecting stationary foreground regions becomes crucial to identify potential objects of interest in many high level applications such as abandoned object detection [1][2]. Stationarity is defined as an object, person or group of people remaining stopped after previous movement. Static foreground detection approaches tend to use Background Subtraction (BS) techniques to detect foreground by comparing frames against a model of the scene's background [3]. They exhibit limitations when operating in crowded environments, adapting to fast illumination changes (causing false detections) and keeping stationary objects as foreground due to their absorption into the background model. Such detection approaches employing background subtraction must develop strategies to handle these issues.

State-of-the-art approaches apply different techniques over foreground data to determine stationarity via binary masks. Temporal foreground accumulation is often used [4][5] to address occlusions. However, it requires efficient handling of illumination changes by the BS model and presents false positive detections when continuous motion occurs, which is common in crowds. Recent approaches combine such accumulation with motion information to deal with high density motion areas [6] and illumination changes [7]. Furthermore, edge analysis can be used to remove wrong detections due to ghost effects (uncovered background due to moving objects that were static during BS training) [8]. Temporal sampling of foreground masks is widely used [9] which can be also combined with motion [10]. Nevertheless, their major drawback lies in the selection of the sampling frequency which has a direct impact on performance. Dual BS approaches [11][12] adapt to slow illumination changes. However, missed and false detections frequently occur in dual BS due to, respectively, undesirable updates of static regions and wrong initialization of the BS model. In [13][14] such limitations are solved by including a finite-state-machine to model pixel history values and using edge information. In [15] several filters (motion, temporal, appearance and edge) are applied to detect stationarity. Finally, some approaches detect stationarity through the relations between the states of BS models based on Mixture-of-Gaussians [16] whilst dealing with ghost detections [17] and illumination changes [18][19]. In summary, state-of-art approaches combine different features to cope with BS limitations in crowds. However, the high rate of false detections limits their use in crowded environments.

We present an approach to detect stationary foreground which accumulates spatio-temporal features to address the aforementioned limitations in crowds. Its contribution is twofold. First, it extends [6] by proposing a generic framework to combine multiple features operating with standard BS models. Second, it introduces a new feature based on Structural Similarity (SSIM) [20] which increases the robustness against illumination changes. Then, a multi-feature stationary detector is created by combining the spatio-temporal evolution of such structural feature with existing foreground and motion features [6]. Unlike other approaches, this proposal faces many challenges in crowds such as occlusions, high dense situations, shadows and illumination changes. Results demonstrate that the structural feature reduces the false positive rate and the performance increase of the multi-feature detector as compared to the state-of-the-art.

The structure of this paper is as follows. Section 2 describes the multi-feature stationary detection framework. Section 3 presents the structural feature and Section 4 briefly overviews the proposed multi-feature approach. Experimental results are discussed in Section 5. Finally, Section 6 summarizes the main conclusions.

2. FRAMEWORK

This section describes the framework to detect stationary foreground via temporal evolution of multiple features, which generalizes the two-feature combination of [6]. Figure 1 presents the proposed framework. For each feature, two common stages take place: *Feature Map* (*FM*) *extraction* and *History Images* (*HI*) *computation*. Then the *Combination & Thresholding* stage combines the feature results to obtain the Stationary Foreground Detection mask.

This framework starts by extracting N Feature Maps $FM_t^{f_i}(\mathbf{x})$ from each frame I_t of the video sequence at time t:

$$FM_t^{f_i}(\mathbf{x}) = g(I_t(\mathbf{x})), \quad (1)$$

where $\mathbf{x}$ is a 2D pixel location, f_i $(i = 1, ..., N)$ are the N features and $g(\cdot)$ describes the process to generate the feature map. Note that

This work has been partially supported by the Spanish Government (TEC2011-25995 EventVideo). We also would like to thank Jose M. Martínez for fruitful discussions.

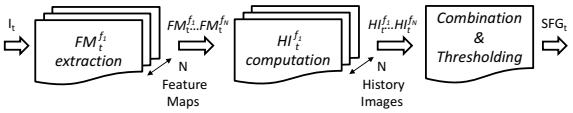

Fig. 1. Overview of the stationary detection framework.

$g(\cdot)$ can represent an extraction process of diverse nature such as a BS approach [21], scene motion [10] or structural information [20].

Then, feature maps are accumulated over time to compute the History Image $HI_t^{f_i}(\mathbf{x})$ of each feature:

$$HI_t^{f_i}(\mathbf{x}) = HI_{t-1}^{f_i}(\mathbf{x}) + \left[w_p^{f_i} \cdot FM_t^{f_i}(\mathbf{x})\right] - \left[w_n^{f_i} \cdot (\sim FM_t^{f_i}(\mathbf{x}))\right], \quad (2)$$

where $w_p^{f_i}$ and $w_n^{f_i}$ are two weights to manage the contribution of $FM_t^{f_i}(\mathbf{x})$ to $HI_t^{f_i}(\mathbf{x})$ for each feature f_i. By default, every $w_p^{f_i}$ and $w_n^{f_i}$ should be set to 1 (or the current value of $HI_{t-1}^{f_i}(\mathbf{x})$) to increase (or reset) the History Image for each feature f_i. Thus, $HI_t^{f_i}(\mathbf{x})$ represents the number of consecutive frames maintaining the desired feature in $FM_t^{f_i}$, which measures stationarity at pixel-level.

Subsequently, History Images are computed for each feature and normalized to the range [0,1], considering the video framerate (25 fps) and the stationary detection time (t_{static} = 20 seconds) to obtain the maximum value when the feature f_i is present for t_{static} consecutive seconds: $\overline{HI_t^{f_i}}(\mathbf{x}) = min\left\{1, HI_t^{f_i}(\mathbf{x}) \,/\, (fps \cdot t_{static})\right\}$.

Then, a Stationary History Image (HI_t^S) is obtained as combination of all $HI_t^{f_i}$ to model their joint stationary variation over time:

$$HI_t^S(\mathbf{x}) = h\left(HI_t^{f_1}(\mathbf{x}), ..., HI_t^{f_N}(\mathbf{x})\right), \quad (3)$$

where $h(\cdot)$ defines the combination rule of the N feature maps such as the mean [6] or sampling [10] rules, which depend on the application and reliability of the features considered.

Finally, the Stationary Detection Mask, $SFG_t(\mathbf{x})$, is computed by thresholding $HI_t^S(\mathbf{x})$:

$$SFG_t(\mathbf{x}) = \begin{cases} 1\ if & HI_t^S(\mathbf{x}) \geq \eta \\ 0 & otherwise \end{cases}, \quad (4)$$

where η should be 1 to trigger detections after t_{static} seconds and $HI_t^S(\mathbf{x}) = 1$ means that every $HI_t^{f_i}(\mathbf{x})$ is stationary for the analyzed pixel $\mathbf{x}$, thus $SFG_t(\mathbf{x})$ is set to 1.

3. STRUCTURE SIMILARITY FEATURE

The main features used for stationary detection (foreground and motion) do not handle illumination changes, limiting their efficiency. This section presents the proposed feature to address such problem.

We propose to use structural information via the Structural Similarity (SSIM) feature [20], originally developed for image quality assessment (i.e., between modified and distortion-free images), returning value 1 for highest quality. SSIM compares two images (a, b) using three components; luminance l, contrast c and structure s:

$$SSIM(a, b) = l(a, b) \cdot c(a, b) \cdot s(a, b), \quad (5)$$

and computes every component over each pixel neighborhood, thus providing a SSIM map at pixel level. We obtain this SSIM map for stationarity detection by comparing the current frame and the background model of [21]. Such comparison determines which pixels belong to object (or background) due to their different (or equal) structure to the background model. Figure 2 shows that SSIM identifies shadows and illumination changes areas as background, having high scores when comparing current frame and background.

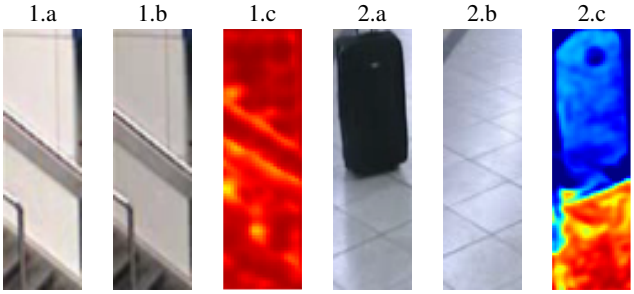

Fig. 2. SSIM map (c) between frame (a) and background (b) patches, where the examples are: (1) an illumination change and (2) an object with its shadow. Dark blue (red) refers to min (max) SSIM scores. In example 1, SSIM (1.c) has high similarity scores despite the different illumination of frame (1.a) and background (1.b). In example 2, SSIM (2.c) has high (low) values in the shadowed (suitcase) area when comparing frame (2.a) and background (2.b).

The proposed structural information feature, Region SSIM map (RSIMM), is the mean of the SSIM map over a square window of size $Q \times Q$ centered at the pixel under analysis (minimum score is bounded to zero). The mean operation is applied to handle the performance decrease of SSIM when after an illumination change, the color is locally saturated. The higher Q the higher the robustness against saturation but the lower the precision. We have selected $Q = 12$ experimentally as a good balance between both saturation and precision. We compute the structural Feature Map as:

$$FM_t^{ST} = \begin{cases} 1\ if & RSSIM_t \leq \tau \\ 0 & otherwise \end{cases}, \quad (6)$$

where τ = 0.5 is a threshold to get a normalized map where $FM_t^{ST} = 1(0)$ when $RSSIM_t \in [0, 1]$ is under (over) τ.

4. MULTI-FEATURE STATIONARY DETECTION

We compose a multi-feature stationary foreground detector by combining foreground, motion and structural information (thus $N = 3$ in Figure 1). The Foreground feature (*FG*) can be extracted via any BS approach. We use [21] for considering pixel neighborhood and noise level to detect foreground. *FG* presents many false positives when continuous motion or BS problems exist. The Motion feature (*MO*) is computed using median filters over temporal sliding windows as described in [6]. Unlike the classical inter-frame difference motion extraction, *MO* handles situations where stationary objects are continuously occluded by high motion, removing false detections while keeping such objects. The proposed structural feature (*ST*) adds robustness against illumination changes and shadows, which mainly affect the foreground map. *ST* compares each new frame against the background model created by [21].

We apply the framework described in Section 2 to obtain three feature maps (FM_t^{FG}, FM_t^{MO} and FM_t^{ST}) and their normalized History Images ($\overline{HI_t^{FG}}$, $\overline{HI_t^{MO}}$ and $\overline{HI_t^{ST}}$) which model

Criteria	Non-crowded					Crowded																Total
	AVSS07	PETS06			PETS07	AVSS07				PETS07		PETS06							HALL			
	Easy	S7_C3	S4_C3	S4_C4	S5_C3	Med	Hard	AB	PV	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	L	H	H	H	M	H	M	H	H	H	M	H	-
Illumination changes	L	-	-	-	M	L	L	L	H	M	M	-	-	-	-	-	-	L	L	L	-
Motion level	L	L	L	L	L	M	H	H	H	H	H	H	M	H	L	H	H	H	M	H	-
Overall complexity	L	L	L	L	L	H	H	H	H	H	H	H	M	H	M	H	H	H	M	H	-
Number of frames	4291	3401	3051	3051	2900	4834	5311	32875	26750	2900	2900	3401	3401	3021	3021	3051	3051	10000	10834	15102	147146
Correct background	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	Yes	-
Annotated stationary regions	2	1	3	4	2	14	13	39	10	3	3	1	1	2	3	6	3	3	2	12	127

Table 1. Description of the sequences of the evaluation set. (Key. L:Low. M:Medium. H:High). Correct background means that a background free of foreground objects is manually captured at the beginning of the sequence.

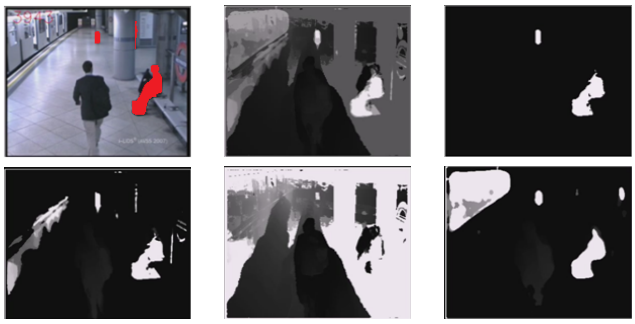

Fig. 3. Example of multi-feature stationary detection example from Hard sequence. From left to right, first row: frame 3943 (ground truth in red), HI^{S}_{3943}, SFG_{3943}; second row: $\overline{HI^{FG}_{3943}}$, $\overline{HI^{MO}_{3943}}$ and $\overline{HI^{ST}_{3943}}$. A missed detection due to a thin stationary region behind the column occurs. $\overline{HI^{FG}_{3943}}$ does not detect the train cars because a selection mask is applied. HI^{S}_{t} shows the combination of all HI (second row), being brighter for stationary regions. The Static Mask (SFG_t) shows the final result (i.e., HI^{S}_{t} thresholding).

foreground-motion-structure variation over time. We generate the final map HI^{S}_{t} as the mean of three History Images. Finally, the Stationary Detection Mask (SFG_t) is computed by thresholding as (slight variation of Eq. (4) to handle $\overline{HI^{MO}_{t}}$ variability):

$$SFG_t = \begin{cases} 1\,if & \overline{HI^{FG}_{t}} \geq \eta \cap \overline{HI^{ST}_{t}} \geq \eta \\ & \cap\, HI^{S}_{t} \geq \eta \times factorTh \\ 0 & otherwise \end{cases} \quad (7)$$

where threshold η (set to 1) maintains the alarm time established with $\overline{HI^{FG}_{t}}$ and $\overline{HI^{ST}_{t}}$ conditions, and *factorTh* weights η over HI^{S}_{t} due to reductions of motion History Image ($\overline{HI^{MO}_{t}}$) over continuously occluded objects where it is difficult to find long periods of the no motion case. As referred in [6], an occlusion handling method is applied to recover lost initial detections by reductions of $\overline{HI^{MO}_{t}}$. An example of the multi-feature detection is given in Figure 3.

5. EXPERIMENTAL RESULTS

This section evaluates the proposed structural feature and compares the multi-feature detector against state-of-the-art approaches.

5.1. Datasets and performance measures

We have evaluated the proposed approach over well-known public datasets (PETS 2006[1], PETS 2007[2] and AVSS 2007[3]) using 17 sequences and constituting a new ground-truth of static regions[4]. The data presents many challenging situations related with multiple occlusions, illumination changes and dense crowds. To extend such complex situations, we have recorded an additional dataset in a faculty hall that introduces a new concept of stationary region, stationary crowd (group of static people in the same spatial location whose size could oscillate over time). The overall evaluation set contains 147146 frames and 127 annotations. Table 1 details the full dataset.

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

$$P = TP/(TP + FP), \quad (8)$$

$$R = TP/(TP + FN), \quad (9)$$

$$F = (2 \cdot P \cdot R)/(P + R), \quad (10)$$

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

Features	**Non-crowded**			**Crowded**		
	P	**R**	**F**	**P**	**R**	**F**
FG	0.73	**1**	0.81	0.33	0.84	0.44
ST	0.73	**1**	0.82	0.47	**0.91**	0.58
FG & MO	0.73	**1**	0.81	0.39	0.83	0.48
ST & MO	0.83	**1**	0.89	0.50	**0.91**	0.60
FG & MO & ST	**0.93**	**1**	**0.96**	**0.53**	0.87	**0.62**

Table 2. Results for combinations of the three features FG, MO and ST to detect stationary foreground. Bold indicates best results.

5.2. Features comparison

Table 2 shows the performance of different combinations for the three features of the proposed approach (*FG*, *MO* and *ST*). In non-crowded scenes *ST* does not improve significantly the results compared with *FG*. However in crowded scenes where many illumination changes and shadows take place, *ST* removes false detections improving the detection accuracy . *MO* information improve results specially in crowded scenes where high density motion areas exist. For combinations, Recall is better for *ST* and *ST & MO* due to a better performance against camouflages than configuration including *FG*, thus missing less detections. Nevertheless, Precision from *ST & MO* is less accurate than *FG & MO & ST* due to worse shapes in their detection masks, leading to more pixels overlapping in History Images computation that increases wrong scores benefiting when stationary objects are removed. In summary, combining *FG, MO* and *ST* removes false detections in high density areas caused by shadows and illumination changes, which *FG* does not handle alone. Some

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

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

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

[4] Ground-truth and software are available at http://www-vpu.eps.uam.es/publications/MFSFD/

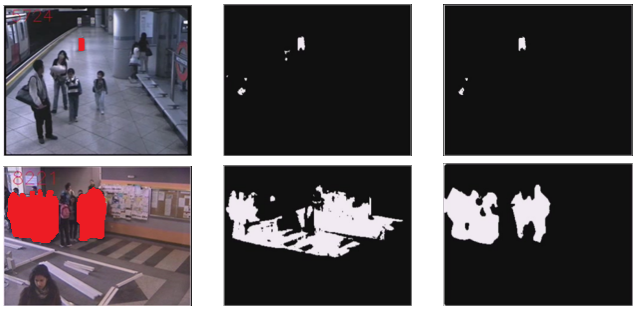

Fig. 4. Comparative examples of feature results. From left to right, first row: Frame 5724 from AB, SFG_{5724}computed with FG and SFG_{5724} computed with *FG & MO*; second row: Frame 8221 from H_S3, SFG_{8221}computed with FG and SFG_{8221}computed with *ST* feature. First row shows how *MO* reduces the false positives due to previous high motion. Second row shows how, unlike *FG*, *ST* is able to tackle illumination changes due to high density in the scene, thus providing more accurate detections free of false positives.

Approach	Non-crowded			Crowded		
	P	R	F	P	R	F
ACC [4]	0.72	**1**	0.80	0.29	**0.87**	0.42
DUAL [11]	0.42	**1**	0.58	0.24	0.76	0.33
SUB [9]	0.67	**1**	0.77	0.25	**0.87**	0.37
BAY [10]	0.74	**1**	0.82	0.34	0.70	0.43
MED [6]	0.73	**1**	0.81	0.39	0.83	0.48
Proposed	**0.93**	**1**	**0.96**	**0.53**	**0.87**	**0.62**

Table 3. Comparative results of the proposed approach against related state-of-the-art. Bold indicates best results.

examples of *MO* and *ST* advantages to remove false detections are shown in Figure 4.

5.3. State-of-the-art comparison

We compare the proposed approach with relevant state-of-the-art based on temporal accumulation of *FG* [4] (ACC) and *FG-MO* [6] (MED), temporal sampling of *FG* [9] (SUB) and *FG-M*otion [10] (BAY) and dual BS [11] (DUAL). ACC, SUB and DUAL have been implemented according to their description whereas the original software is used for MED and BAY. All approaches use the default settings proposed by their respective authors.

Table 3 shows how the selected approach achieves higher performance than selected state-of-the-art, especially in crowds, where an increase of around 29% in F-score is obtained. Such enhancement is due to the high reduction of false detections (increasing P) in cases of shadows and illumination changes. The Recall value is maintained as best state-of-the-art approaches as the same regions are detected. Figure 5 shows three visual examples where the improvement against related work approaches can be appreciated.

6. CONCLUSIONS

This paper formalizes a multi-feature framework for stationary foreground detection in complex scenarios. Fast illumination changes are handled by a structural comparison between frame and background model. Adding structural information with foreground and

Fig. 5. Comparative results of selected approaches. From top to bottom: Frame and SFG_t from ACC, SUB, BAY, DUAL, MED and proposed approach. From left to right, frames 3993, 9081 and 7423 from Hard, PV and H_S3 sequences respectively. Except the proposed method, no one handles shadows and illumination changes. Although BAY uses motion information for temporal sampling, the random sampling nature together with the issues of inter-frame difference result in missing detections when continuous motion takes place. MED solves motion issues, however shadows and sudden illumination changes end in many false detections. The proposed approach addresses such problems, having the initialization of the background model as its main source of error.

motion features provides a suitable approach to operate in crowded environments due to its robustness against occlusions, shadows, illumination changes and high density motion situations. Experimental results show a notable performance improvement over state-of-art approaches in challenging datasets. The main drawback of the proposed approach is the dependency from the initial instants to capture a proper background, decreasing the performance when a scene free of foreground objects is not available due to uncovered background situations. Future work will investigate efficient background initialization and the use of edge information to avoid ghost detections.

7. REFERENCES

[1] C. Beleznai, P. Gemeiner, and C. Zinner, "Reliable left luggage detection using stereo depth and intensity cues," in *Proc. of*

IEEE Conf. Computer Vision (ICCV) Workshops, Jun. 2013, pp. 59–66.

[2] A. Lopez-Mendez, F. Monay, and J.M. Odobez, "Exploiting scene cues for dropped object detection," in *Proc. of Int. Joint Conf. Computer Vision, Imaging and Computer Graphics Theory and Applic. (VISAPP)*, 2014, pp. 1–9.

[3] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino, "Background subtraction for automated multisensor surveillance: A comprehensive review," *EURASIP J. Adv. Signal Process*, vol. Article ID 343057, pp. 1–24, Feb. 2010.

[4] S. Guler and J. A. Silverstein, "Stationary objects in multiple object tracking," in *Proc. of IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sept. 2007, pp. 248–253.

[5] L. Maddalena and A. Petrosino, "Stopped object detection by learning foreground model in videos," *IEEE Trans. Neural Net. and Learning Syst.*, vol. 24, no. 5, pp. 723–735, May 2013.

[6] D. Ortego and J.C. SanMiguel, "Stationary foreground detection for video-surveillance based on foreground and motion history images," in *Proc. of IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2013, pp. 75–80.

[7] W. Hassan, P. Birch, B. Mitra, N. Bangalore, R. Young, and C. Chatwin, "Illumination invariant stationary object detection," *IET Computer Vision*, vol. 7, no. 1, pp. 1–8, Feb. 2013.

[8] J. Pan, Q. Fan, and S. Pankanti, "Robust abandoned object detection using region-level analysis," in *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Sept. 2011, pp. 3597–3600.

[9] C. Jing-Ying, L. Huei-Hung, and C. Liang-Gee, "Localized detection of abandoned luggage," *EURASIP J. Adv. Signal Process.*, vol. Article ID 675784, pp. 1–10, 2010.

[10] 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)*, Sept. 2010, pp. 4657–4660.

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

[12] 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.*, vol. 7, no. 1, May 2013.

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

[14] R.H. Evangelio and T. Sikora, "Complementary background models for the detection of static and moving objects in crowded environments," in *Proc. of IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2011, pp. 71–76.

[15] 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)*, Sept. 2012, pp. 70–75.

[16] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. of IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, Jun. 1999, vol. 2, pp. 246–252.

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

[18] Q. Fan and S. Pankanti, "Robust foreground and abandonment analysis for large-scale abandoned object detection in complex surveillance videos," in *Proc. of IEEE Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sept. 2012, pp. 58–63.

[19] Y. Tian, A. Senior, and M. Lu, "Robust and efficient foreground analysis in complex surveillance videos," *Mach. Vision Appl.*, vol. 23, no. 5, pp. 967–983, Sept. 2012.

[20] H. R. Sheikh Z. Wang, A. C. Bovik and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[21] A. Cavallaro, Steiger O., and Ebrahimi T., "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200–1209, Oct. 2005.