Bio-Inspired Boosting for Moving Objects Segmentation

Isabel Martins[1,2✉], Pedro Carvalho[2,3], Luís Corte-Real[3,4], and José Luis Alba-Castro[1]

[1] University of Vigo, Vigo, Spain
[2] School of Engineering, Polytechnic Institute of Porto, Porto, Portugal
mis@isep.ipp.pt
[3] INESC TEC, Portugal
[4] Faculty of Engineering, University of Porto, Porto, Portugal

Abstract. Developing robust and universal methods for unsupervised segmentation of moving objects in video sequences has proved to be a hard and challenging task. State-of-the-art methods show good performance in a wide range of situations, but systematically fail when facing more challenging scenarios. Lately, a number of image processing modules inspired in biological models of the human visual system have been explored in different areas of application. This paper proposes a bio-inspired boosting method to address the problem of unsupervised segmentation of moving objects in video that shows the ability to overcome some of the limitations of widely used state-of-the-art methods. An exhaustive set of experiments was conducted and a detailed analysis of the results, using different metrics, revealed that this boosting is more significant when challenging scenarios are faced and state-of-the-art methods tend to fail.

Keywords: Bio-inspired motion detection · Video segmentation

1 Introduction

Segmentation of moving objects in video sequences is a fundamental step in many computer vision applications. Therefore, the identification of changing or moving areas in a video is a crucial step. Despite the large number of methods proposed in the literature to address the unsupervised segmentation of moving objects, none has been able to fully deal with complex and challenging scenarios that include poor lighting conditions, sudden illumination changes, shadows and parasitic background motion.

Comprehensive reviews of background subtraction (BS) approaches have been presented in [1,2]. Although they provide an overview of existing methods, the results reported by different authors have not been computed on a common dataset, making it hard to establish fair comparisons. Also, many datasets do not contain a balanced set of videos presenting real application challenges. Moreover, metrics used to evaluate the average algorithms' performance do not reveal how they perform frame by frame. Recent research has shown that methods appear to be complementary in nature, with the best-performing methods being beaten by combining several of them [3].

Recently, a considerable number of image processing modules inspired in biological models of the human visual system have been explored [4,5,6]. The ultimate goal is to copy the recognition capability of the human visual system. The image processing occurring at the level of the human retina allows not only noise and illumination variation removal, but also static and dynamic contours enhancement. Hence, this approach can be used for illumination normalization and motion detection.

This paper proposes a new scheme to address the problem of unsupervised segmentation of moving objects, which exploits the fusion of information obtained from two inherently different approaches: a bio-inspired motion detection method, using low-level information from the modeling of the human visual system, and a BS algorithm based on pixel color information. The biologically inspired model of the human retina presented in [6] has been adopted for the former. Experiments were performed with several BS algorithms showing that our method consistently improves the results, particularly in complex situations, where the BS algorithms critically fail.

The paper is organized as follows. Section 2 introduces the bio-inspired model of the retina that motivated our proposal. Section 3 presents the bio-inspired motion segmentation method. The experimental setup and the obtained results are presented in sections 4 and 5, respectively. Final conclusions are presented in section 6.

2 The Retina Model

Figure 1 presents the global architecture of the adopted retina model [6] as a combination of low-level processing modules. Basically, it is a layered model with: 1) photoreceptors, where local contrast is enhanced; 2) outer plexiform layer (OPL), where the non-separable spatio-temporal filtering removes spatio-temporal noise and enhances spatial high-frequency contours while reducing or removing the mean luminance; 3) inner plexiform layer (IPL), with two channels: the parvocellular (Parvo) channel, dedicated to spatial analysis enhancing static contours contrast, and the magnocellular (Magno) channel that enhances moving contours and removes static ones.

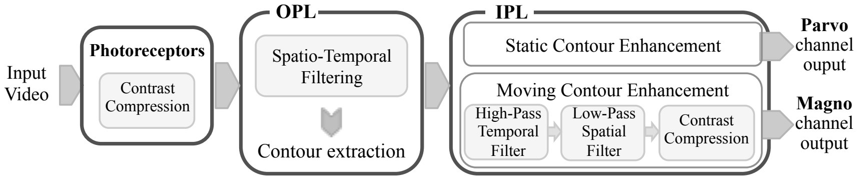

Fig. 1. The retina model proposed in [6].

As our goal is to extract the regions with moving contours, we will focus on the Magno channel output. A temporal effect is introduced on its output signal. This effect is modeled by a first order high-pass temporal filter, with transfer function given by (1), where τ is the temporal constant of the filter. This filter enhances changed areas. Its output is smoothed by a spatial low-pass filter. Finally, local contrast compression enhances the resulting contour information.

$$A(z) = b\frac{1 - z^{-1}}{1 - bz^{-1}} \quad \text{with} \quad b = e^{-1/\tau} \tag{1}$$

The Magno channel output signal magnitude is dependent on the velocity of the moving areas, with high response for fast moving areas and null response for static regions. The response of the filter is also stronger for moving contours perpendicular to the motion direction. The tuning of the temporal constant allows the adjustment of the response to temporal changes in the scene. A low value allows the enhancement of only fast changes whereas a higher value allows the enhancement of slower changes. It affects not only the response to the contours of moving objects, but also to parasitic background motion. The response decays with time leading to fuzzy contours.

3 Bio-Inspired Hybrid Segmentation Method

The proposed bio-inspired hybrid segmentation, represented in Figure 2, merges information from two inherently different approaches: 1) the bio-inspired motion segmentation that identifies regions of motion; 2) a BS method, based on pixel color information, that extracts the silhouettes of the moving objects. The final foreground mask, called *HybridMask,* is obtained by merging the outputs of the two modules.

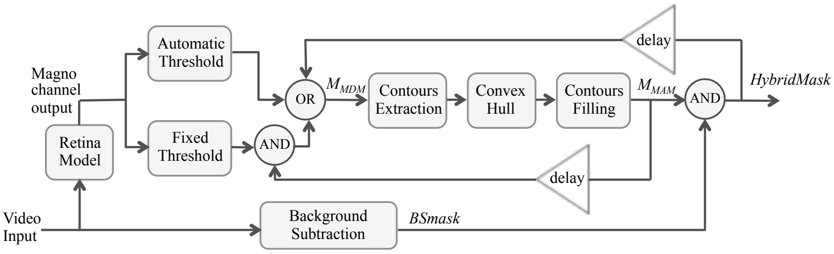

Fig. 2. Block diagram of the bio-inspired hybrid segmentation method

The bio-inspired motion segmentation consists of the segmentation of the Magno channel output signal of the retina model, which allows the detection of transient events (motion, changes) with reduced noise errors even in difficult lighting conditions. It is a low spatial frequency signal that gives a coarse representation of contours enabling motion blobs to be reliably extracted.

The high-pass temporal filter of the Magno channel introduces a temporal effect on the output signal, clearly visible as a trace left by the moving objects. This parameter can be set up in the configuration of the model, allowing tuning of the retina model to the characteristics of the input video sequence. However, in all the experiments reported, the default value (2.0) was used. The variable delay introduced by the filter depends on the value of its temporal constant. Hence, if the range of the apparent velocities of the objects is known in advance, a delay can be introduced in the BS module to compensate for this delay. However, experimental results have shown that, with the temporal constant used, a null delay provides good results.

The Magno segmentation process leads to the definition of the Magno Moving Areas Mask, M_{MAM}. The process can be summarized as follows. Let M be the Magno channel output signal. In the absence of moving objects, its magnitude is very low, corresponding to residual noise. In the presence of moving objects, the magnitude of M takes higher values in the neighborhood of moving contours. First, a Magno Motion Detection Mask, M_{MDM}, is created according to:

$$M_{MDM}(x,y,t) = \begin{cases} 1, & if \ \ (M(x,y,t) > T_{Var} \ \ or \ \ HybridMask(x,y,t-1) = 1 \\ & \quad or \ \ (M(x,y,t) > T_{Low} \ \ and \ \ M_{MAM}(x,y,t-1) = 1)) \\ 0, & otherwise \end{cases} \quad (2)$$

T_{Var} is a dynamically changing global threshold, taking the value of 2.5 standard deviations of the magnitude of the Magno channel output signal. To get temporally stable regions of motion, the M_{MDM} is connected to the previous M_{MAM} by adding pixels with a value above a fixed value T_{Low} (experimentally set to 15.0) that were set in the previous M_{MAM}. The Magno segmentation fails to detect still foreground objects as it relies on motion information. To avoid loosing stopped foreground objects, pixels that were set in the previous *HybridMask* are also added to the M_{MDM}.

Finally, to create the M_{MAM} from the M_{MDM}, a connected component analysis is performed to extract the exterior closed contours. To avoid highly non-convex blobs, the convex hull of these contours is calculated using the algorithm presented in [8]. The final M_{MAM} is obtained by filling the contours. This mask contains the regions of motion where the objects silhouettes are to be extracted by the BS algorithm.

Figure 3 shows an example of the Magno channel output and the resulting M_{MAM}.

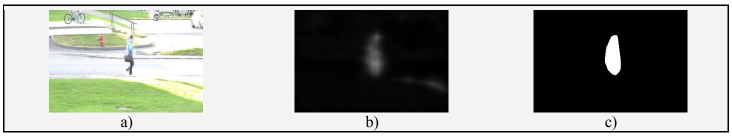

Fig. 3. a) Input frame, b) Magno channel output (M), c) Magno Moving Areas Mask (M_{MAM})

As stated before, the Magno channel output signal gives a fuzzy representation of contours. This allows the extraction of temporally stable motion blobs, but not the precise contours of the objects. Thus, a pixel-level BS algorithm based on pixel color information is needed to extract these contours. Experiments were performed with several state-of-the-art BS algorithms, reported in [3]. However, other algorithms could be used in this module. The fact that GMM is widely used, and that finding the best parameter set for a particular application is not a trivial task, often leads to the use of the default parameters. For this reason, we decided to include it in our experiments using those settings. For each input frame, the foreground mask resulting from the BS algorithm is referred to as *BSmask*.

The fusion step combines the complementary information resulting from the two approaches to enhance overall detection accuracy. The Magno segmentation produces spatially and temporally coherent regions due to the spatio-temporal integration performed by the retina. These results are robust to spatio-temporal noise, global illumination changes and soft shadows, but the masks tend to be larger than the objects due to the fuzziness of the contours. On the other hand, BS algorithms perform well in extracting the silhouettes of foreground objects in a large number of situations, but are less robust. Their performance is also highly dependent on the correct tuning of the parameters. The fusion step uses the regions provided by the Magno channel segmentation to focus the foreground detection. The final foreground mask, *HybridMask*, is created according to:

$$HybridMask(x,y,t) = \begin{cases} 1, & if \quad M_{MAM}(x,y,t) = 1 \quad and \quad BS_{mask}(x,y,t) = 1 \\ 0, & otherwise \end{cases} \tag{3}$$

4 Experimental Setup

An exhaustive set of experiments was conducted to evaluate the performance of the proposed method compared with the base BS method. Only one set of parameters was used for all the videos. The default parameter set was used for the setup of the retina model (available in OpenCV). The bio-inspired motion segmentation module is running with no configurable parameters. For evaluation purposes, several alternatives were used as BS method: MOG2, refers to the masks outputted by MOG2, available in OpenCV, using default parameters; GMM [7], KNN [7], AMBER [11], CwisarDH [12], Spectral360 [13], SuBSENSE [14] and FTSG [15] refer to the computed masks made available in the CDnet site [9]. These masks were generated with the parameters adjusted to maximize overall performance.

The experiments were conducted on the complete set of videos of the CDnet 2014 Dataset [9]. Evaluation was performed using the ground truth (GT) segmentation provided along with the videos. Each mask can have 5 labels: *Moving*, corresponding to foreground pixels; *Static*, corresponding to background pixels; *Shadow* corresponding to moving shadows; *Non-ROI* corresponding to regions outside the ROI; *Unknown* corresponding to pixels whose status is unclear.

The following seven metrics are often used to rank BS methods [3] [10]: Recall (Re), Specificity (Sp), False Positive Rate (FPR), False Negative Rate (FNR), Percentage of Wrong Classifications (PWC), Precision (Pr) and F-measure. We assessed the proposed and base methods over each video by computing these metrics, followed by a category-average and an overall-average metric. In our comparisons, the F-measure was used as an indicator of performance since, as reported in [3] [10], it correlates most strongly with the rankings produced by evaluation algorithms.

Considering the image segmentation as a partition, a metric based on the normalized symmetric distance between partitions, d_{sym}, was proposed in [16]. This metric has shown to be consistent with the subjective evaluation that a human observer would make and can provide an error value for each of the frames.

These complementary metrics allowed us to evaluate the improvement achieved by the proposed method, and identify failures. When computing the metrics, pixels classified as *Shadow* are considered as *Static* and pixels classified as *Non-ROI* or *Unknown* are discarded.

5 Analysis of Results and Discussion

Table 1 shows the average values for the first set of metrics across all categories for the overall set of videos. The proposed bio-inspired boosting method consistently outperforms the base method. As expected, as the base BS algorithm quality improves, the boosting achieved by the fusion with the bio-inspired motion segmentation is lower. However, for the eight algorithms tested, the overall measures improve using the hybrid method even if, for some categories, there is some marginal decrease in performance. There are also some scenarios where we should not expect to achieve improvements with the proposed method, like in the intermittent object motion category. Table 2 shows the average F-measure for each category. Mind that the best methods are complex algorithms that already combine different approaches.

Table 1. Overall results across all categories.

Method	Re	Sp	FPR	FNR	PWC	Pr	F-measure
MOG2	0.535	0.979	0.021	0.464	3.836	0.508	0.430
Hybrid-MOG2	**0.542**	**0.972**	**0.011**	**0.457**	**2.910**	**0.670**	**0.515**
GMM	0.660	0.971	0.028	0.339	4.052	0.611	0.568
Hybrid-GMM	**0.669**	**0.972**	**0.028**	**0.331**	**4.026**	**0.623**	**0.579**
KNN	0.662	0.980	0.020	0.338	3.363	0.675	0.596
Hybrid-KNN	**0.670**	**0.980**	**0.019**	**0.329**	**3.314**	**0.687**	**0.607**
AMBER	0.722	0.963	0.020	0.278	2.808	0.712	0.666
Hybrid-AMBER	**0.720**	**0.965**	**0.018**	**0.279**	**2.647**	**0.724**	**0.673**
Spectral360	0.748	0.951	0.015	0.252	2.370	0.718	0.690
Hybrid- Spectral360	**0.741**	**0.952**	**0.014**	**0.258**	**2.283**	**0.729**	**0.694**
CwisarDH	0.681	0.977	0.006	0.319	1.536	0.775	0.706
Hybrid-CwisarDH	**0.687**	**0.978**	**0.005**	**0.312**	**1.475**	**0.787**	**0.742**
SuBSENSE	0.806	0.974	0.009	0.194	1.663	0.752	0.742
Hybrid-SuBSENSE	**0.802**	**0.975**	**0.008**	**0.198**	**1.615**	**0.759**	**0.745**
FTSG	0.786	0.975	0.007	0.214	1.272	0.775	0.746
Hybrid-FTSG	**0.785**	**0.976**	**0.007**	**0.215**	**1.247**	**0.781**	**0.750**

Table 2. Average % of improvement in F-measure for each category and across all categories.

Category	MOG2	GMM	KNN	AMBER	Spectral360	CwisarDH	SuBSENSE	FTSG
badWeather	16.09	1.71	1.67	-0.11	-0.25	0.48	0.06	-0.04
baseline	11.23	1.10	0.79	0.50	-0.04	0.72	-0.11	-0.11
cameraJitter	20.87	0.64	0.73	0.04	-0.23	0.63	0.05	0.13
dynamicBackground	89.64	2.11	1.60	0.18	-0.20	1.20	0.04	-0.09
intermittentObjectMotion	-2.60	0.20	-0.71	0.06	-0.41	0.55	-0.85	0.61
lowFrameRate	5.69	1.17	1.88	1.46	-0.01	2.81	-0.11	1.46
nightVideos	18.95	8.30	8.48	11.39	8.12	7.74	6.09	2.55
PTZ	17.49	2.76	2.98	1.56	3.45	4.04	0.28	1.69
shadow	9.43	1.70	2.47	0.63	0.27	0.74	-0.09	0.04
thermal	0.26	1.08	0.82	0.12	-0.32	0.75	0.23	0.32
turbulence	84.59	0.56	0.59	0.02	0.04	0.00	0.03	0.08
Overall	**19.79**	**1.76**	**1.76**	**0.97**	**0.64**	**1.40**	**0.39**	**0.45**

The evaluation of the results using the partition distance metric, d_{sym}, computed frame by frame, gives us a new insight about the performance of the bio-inspired method compared to the base ones. Unlike the F-measure, this is an error measure and, therefore, a lower value means higher quality. To illustrate, Table 3 shows the average F-measure and average d_{sym} for the video *streetCornerAtNight* from the *Night Videos* category, one of the most difficult categories [3]. This video consists of traffic scenes captured at night and the main challenge is to deal with low-visibility of vehicles and their very strong headlights that cause halos and reflections on the street. The learning of the background and foreground detection by the BS methods critically fail in these scenes. However, the retina model processing acts on the input frame for illumination normalization, strongly attenuating variations of illumination. Figure 4, on the left, shows the evolution of the d_{sym} metric from frame 800 to frame 2999 for different base algorithms alone and with boosting. As illustrated, all the algorithms tend to fail in the same frames, corresponding to the most difficult situations, and in these frames the bio-inspired segmentation achieves a significant improvement in the quality of the segmentation. Figure 4, on the right, illustrates the evolution of the F-measure from frame 956 to frame 998, where the d_{sym} shows the first peak (shadowed region). It is clear that the d_{sym} results are consistent with the F-measure results.

Table 3. Average F-measure and average d_{sym} for video *streetCornerAtNight*.

Frame #	800-2999	**MOG2**	**GMM**	**KNN**	**AMBER**	**Spectral360**	**CwisarDH**	**SuBSENSE**	**FTSG**
F-measure	***BS* only**	0.174	0.336	0.349	0.395	0.445	0.350	0.604	0.568
	Hybrid	**0.306**	**0.453**	**0.447**	**0.465**	**0.550**	**0.481**	**0.618**	**0.606**
d_{sym}	***BS* only**	0.034	0.021	0.020	0.018	0.019	0.021	0.013	0.014
	Hybrid	**0.021**	**0.016**	**0.016**	**0.015**	**0.015**	**0.016**	**0.012**	**0.013**

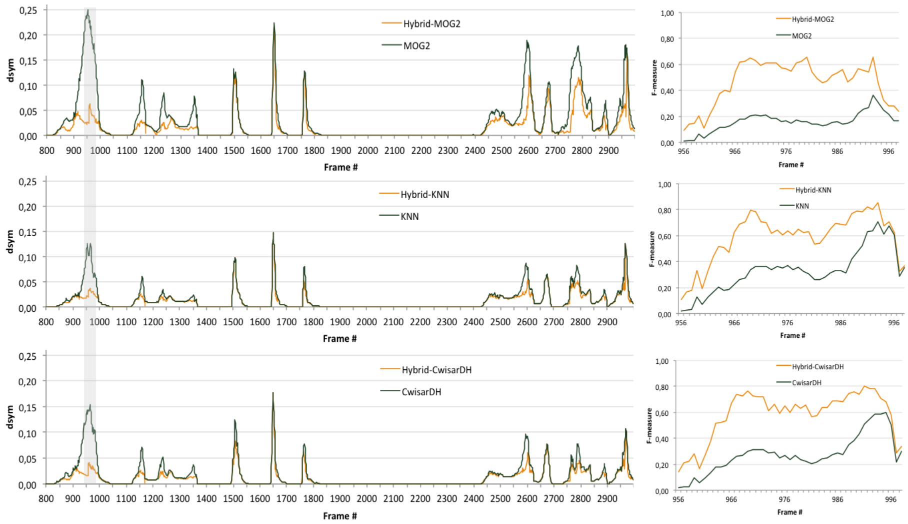

Fig. 4. Left: Evolution of d_{sym} from frame 800 to frame 2999. Right: Evolution of F-measure from frame 956 to frame 998 (shadowed region) of video *streetCornerAtNight*.

Table 4 reports some of the results obtained for the video *copyMachine* from the *Shadow* category, an indoor scene with very noticeable shadows. The hybrid segmentation shows to be much less sensitive to shadows, failing only on very hard shadows.

Figure 5 shows the original frame, the GT, the *BSmasks* and the *HybridMasks* for frame 968 of the video *streetCornerAtNight*, on the left, and frames 2778 and 2816 of the video *copyMachine*, on the right.

Table 4. Average F-measure and average d_{sym} for video *copyMachine*.

Frame # 500-3399	**MOG2**	**Hybrid-MOG2**	**KNN**	**Hybrid-KNN**	**AMBER**	**Hybrid-AMBER**	**CwisarDH**	**Hybrid-CwisarDH**
F-measure	0.506	**0.522**	0.623	**0.653**	0.658	**0.678**	0.878	**0.895**
d_{sym}	0.062	**0.058**	0.056	**0.051**	0.054	**0.050**	0.032	**0.029**

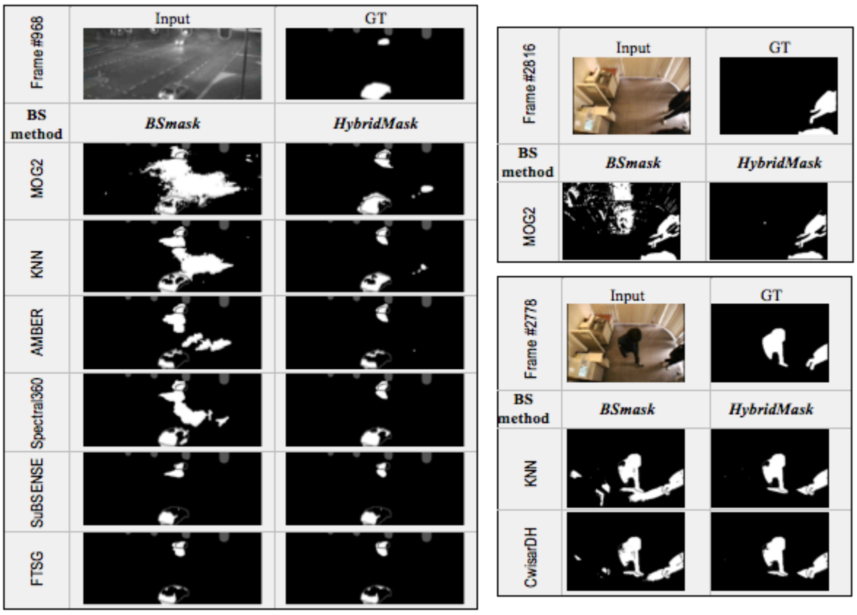

Fig. 5. Foreground masks. Left: video *streetCornerAtNight*. Right: video *copyMachine*.

6 Conclusions

This paper proposes a bio-inspired hybrid method for the unsupervised segmentation of moving objects in video sequences. The proposed method improves well-known and widely used state-of-the-art algorithms in complex situations where these fail. The fusion of the BS method with the proposed bio-inspired motion segmentation greatly reduces the number of false positives. Hence, the combination of the two approaches boosts overall detection accuracy. A detailed analysis of the results, using complementary types of metrics, has revealed that these improvements are more significant when the BS method faces more difficult scenarios, like challenging illumination conditions or shadows, and fails. It must be highlighted that all the experiments in all the testing scenarios were run with the same set of parameters. After the detailed analysis of the results and the identification of the scenarios where the proposed method significantly boosts the segmentation, future work will concentrate in finding the best set of parameters to apply in each situation and automatically adjust them in real time.

Acknowledgements. Work supported by the Galician Regional Government under agreement for funding the Atlantic Research Center for Information and Communication Technologies (AtlantTIC) and research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and project "NORTE-01-0145-FEDER-000020", financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

1. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: An overview. Computer Science Review, May (2014)
2. Elhabian, S. Y., El-Sayed, K. M., Ahmed, S. H.: Moving object detection in spatial domain using background removal techniques — State-of-art. Recent Patents on Computer Science, vol. 1, pp. 32–54 (2008)
3. Wang, Y., Jodoin, P.-M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: CDnet 2014: An Expanded Change Detection Benchmark Dataset. In Proc. IEEE Workshop on Change Detection (CDW-2014) at CVPRW-2014, pp. 387–394 (2014)
4. Itti, L., Koch, C., Niebur,E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Tran. Pattern Anal. Mach. Intell., vol. 20, no. 11, 1254–1259 (1998)
5. Reinhard, E., Devlin, K.: Dynamic range reduction inspired by photoreceptor physiology. IEEE Trans. Visual. Comput. Graphics, vol. 11, pp. 13–24 (2005)
6. Benoit, A., Caplier, A., Durette, B., Herault, J.: Using human visual system modeling for bio-inspired low level image processing. Comput. Vis. Image Underst., vol. 114, no. 7, pp. 758–773 (2010)
7. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Patt. Recogn. Lett., vol. 27, no. 7, pp. 773–780 (2006)
8. Sklansky, J.: Finding the convex hull of a simple polygon. Patt. Recogn. Lett., vol. 1, 79–83 (1982)
9. ChangeDetection.NET (CDNET), `http://www.changedetection.net`
10. Goyette, N., Jodoin, P., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: A new change detection benchmark dataset. In Proc. IEEE Workshop on Change Detection (CDW-2012), at CVPRW-2012 (2012)
11. Wang, B., Dudek, P.: A fast self-tuning background subtraction algorithm. In Proc. IEEE Workshop on Change Detection (CDW-2014), at CVPRW-2014, (2014)
12. Gregorio, M. D., Giordano,M.: Change detection with weightless neural networks. In Proc. IEEE Workshop on Change Detection (CDW-2014), at CVPRW-2014, (2014)
13. Sedky, M., Moniri, M., Chibelushi, C. C.: Spectral-360: A physical-based technique for change detection. In Proc. IEEE Workshop on Change Detection (CDW-2014), at CVPRW-2014, (2014)
14. St-Charles, G.-A. B. P.-L., Bergevin, R.: Flexible background subtraction with self-balanced local sensitivity. In Proc. IEEE Workshop on Change Detection (CDW-2014), at CVPRW-2014, (2014)
15. Wang, G. S. R., Bunyak, F., Palaniappan, K.: Static and moving object detection using flux tensor with split gaussian models. In Proc. IEEE Workshop on Change Detection (CDW-2014), at CVPRW-2014, (2014)
16. Cardoso, J. S., Corte-Real, L.: Toward a generic evaluation of image segmentation. IEEE Trans. on Image Processing, vol. 14, no. 11, pp. 1773–1782 (2005)