Saliency Detection Using Joint Temporal and Spatial Decorrelation

Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä

Center for Machine Vision Research, University of Oulu, Finland
{hamed.rezazadegan,esa.rahtu,janne.heikkila}@ee.oulu.fi
http://www.cse.oulu.fi/CMV

Abstract. This article presents a scene-driven (i.e. bottom-up) visual saliency detection technique for videos. The proposed method utilizes non-negative matrix factorization (NMF) to replicate neural responses of primary visual cortex neurons in spatial domain. In temporal domain, principal component analysis (PCA) was applied to imitate the effect of stimulus change experience during neural adaptation phenomena. We apply the proposed saliency model to background subtraction problem. The proposed method does not rely on any background model and is purely unsupervised. In experimental results, it will be shown that the proposed method competes well with some of the state-of-the-art background subtraction techniques especially in dynamic scenes.

1 Introduction

In recent years, there have been much effort to investigate and develop biological inspired vision systems. Replicating visual attention mechanism is a good example of a bio-inspired vision system. Visual attention can be considered as the process of selectively concentrating on some elements of a visual scene while ignoring others. In computer vision, visual attention models can be used in the preliminary process of localizing eminent information (i.e. regions of interest) from a scene. This process can be referred as detection of saliency.

Saliency detection methods are categorized into top-down and bottom-up approaches [1]. In top-down techniques, usually task-driven factors affect the salience computing process. In the case of bottom-up methods, scene-driven factors are used to distinct perceptual quality which makes some items stand out from their neighbourhood (i.e. visually salient regions).

Saliency detection can improve computer vision algorithms' efficiency by curtailing the processed data. It is applicable to a wide range of applications such as object detection [2], object recognition [3,4], image segmentation [5,6], target tracking [7], image and video compression [8], video retargeting [9], video frame rate conversion [10], image thumbnailing [11,12] and etc. In this paper, we will define a bottom-up saliency model for videos and apply it to the background subtraction problem.

Background subtraction is the overture to many computer vision algorithms; its main objective is localizing objects of interest (e.g. all the moving objects in a

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 707–717, 2013.

traffic monitoring application). To this end, many methods build a background model which describes background and use it to differ background from objects of interest. One major challenge in background subtraction is dynamic of scenes (e.g. camera motion, illumination change), which affects the background model and makes discrimination of objects of interest difficult. In order to deal with this problem and its difficulties many background subtraction techniques have been developed. Bouwmans [13] categories them into basic, statistical, fuzzy, neural network, wavelet background modeling techniques, background clustering and background estimation.

An example of basic background modeling is temporal median filtering which assumes temporal median value of a pixel in a video buffer models background for that pixel [14]. Temporal averaging is another example where the background estimate is computed by recursive update of the average of the history of pixel values [15]. Statistical background modeling techniques utilize statistical analysis to model background pixels; for instance the method of Shauffer and Grimson [16] builds the background model using a mixture of Gaussian. Later, Zivkovic and van der Heijden [17] extended the mixture of Gaussian method to adapt a number of Gaussian mixtures automatically; improving the performance of the algorithm. In order to deal with dynamic background and suppress jitter and noise Elgamnel et al. [18] applied kernel density estimation to approximate probability density function of each pixel by recursive sampling of intensity values taken consequently in a temporal window.

The background subtraction problem can also be solved by saliency detection techniques, if we assume that the object of interest is salient (i.e. differs from its surrounding). Major benefit of applying saliency detection to background subtraction is that it requires no background model. For instance, Mahadevan et al. [19] propose a saliency detection techniques based on center-surround hypothesis [20] and successfully apply it to background subtraction problem. However, their method does not run in real-time. Itti and Baldi [21] exploit differences in posterior and prior beliefs to define salient regions. Their method is not designed specifically for background subtraction, though can be used for such a purpose. In the rest of this paper, we describe a spatio-temporal saliency detection method and explain its application to the background subtraction problem. The proposed saliency detection technique utilizes natural image statistics, runs in real time and competes with the state-of-the-art techniques in background subtraction.

2 Saliency Measure

We propose a purely scene-driven (i.e. bottom-up) saliency measure for videos. In spatial domain, the method consists of several spatial conspicuousness maps (i.e. intermediate saliency maps derived from different feature cues). Each map represents amount of saliency in a frame. To compute conspicuousness maps, we project image patches using basis vectors obtained by non-negative matrix factorization. Also, we compute several temporal conspicuousness maps using

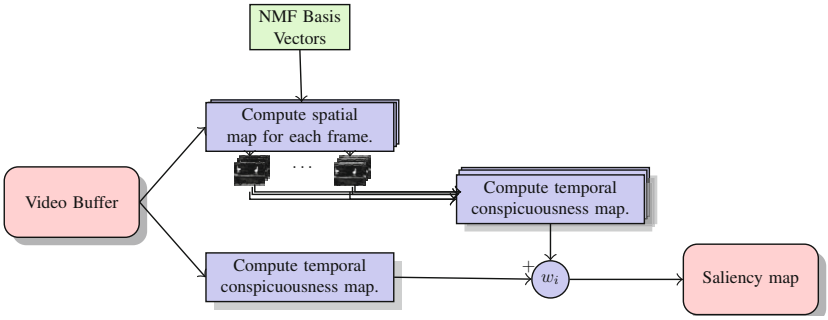


Fig. 1. Overall overview of the proposed framework

principal component analysis. All of these maps are merged together to make a final spatio-temporal map. Figure 1 represents the whole proposed saliency framework. Each part is described with more details in the following sections.

2.1 Spatial Conspicuousness Map

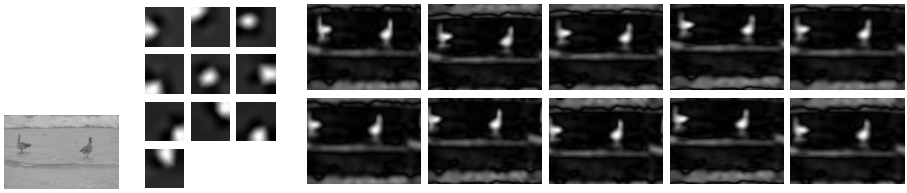
It is demonstrated that learning a sparse code from natural image statistics and projecting image patches using them can produce results similar to those appearing in the primates' primary visual cortex (V1) [22]. A group of saliency techniques rely on sparse coding techniques; for instance, Bruce and Tsotsos [23] applied independent component analysis (ICA) in an information theoretic framework to measure saliency as a function of entropy for still images. Zhang et al. [24] apply a similar formulation in a Bayesian framework to cope with images.

To produce spatial conspicuousness maps, we project image patches using basis vectors obtained by non-negative matrix factorization (NMF). NMF can substitute ICA and approximate neural responses by learning from natural image statistics [25]. Moreover, it provides sparse representation that models neural receptive field responses [26] and helps encoding data with few active elements [27].

The main motivation for using NMF basis over ICA basis vectors is that NMF provides a part-based representation while ICA will result in a holistic image representation [28]. We expect that an intermediate representation (i.e. part-based) performs better as it encodes both local and global information. We test the idea by comparing the proposed method replacing NMF with ICA and PCA in spatial domain.

Computing Conspicuousness Map. Let us assume that we have a video frame I and a set of NMF basis vectors S_i, a spatial conspicuousness map is computed as follows:

$$C_s^i(I) = |S_i * [(I - \mu_I)/\sigma_I]\,|, \tag{1}$$



(a) (b) (c)

Fig. 2. a) An arbitrary image, (b) NMF basis vectors, (c) Corresponding spatial conspicuousness map for the given image

where μ_I is the mean and σ_I is the standard deviation of the frame I and $*$ is the convolution operator. $S_i = w_i, \quad i = 1 \dots r$, w_i is the i_{th} row of $W^\dagger$ (i.e. pseudo-inverse of W) in matrix form[1].

In order to compute W, we use McGill color image dataset [29]. Initially, images are converted to gray-scale and one million patches of size 24×24 were sampled. Each patch is treated as a vector and stacked in column order to make a $n \times m$ matrix $\mathbf{V} = [v_1, v_2, \dots, v_m]$. The following approximate factorization can be written

$$\mathbf{V}_{ij} \approx (WH)_{ij} = \sum_{k=1}^{r} w_{ik} h_{kj}. \tag{2}$$

where each column of W represents the so called basis vector, and each column of H consists of encoding coefficients which define the strength of each basis vector. Having $\mathbf{V}$, we estimate W and H through the following optimization

$$\begin{cases} \underset{W,H}{\text{minimize}} & f(W,H) = \frac{1}{2}\left\|\mathbf{V} - WH\right\|_F^2 \\ \text{subject to} & W, H > 0. \end{cases} \tag{3}$$

where $\|.\|_F$ is the Frobenius norm. Figure 2 depicts 10 basis vectors and corresponding conspicuousness maps obtained using the aforementioned method.

2.2 Temporal Conspicuousness Map

We define a temporal conspicuousness map as a mean to describe stimulus change during neural adaptation phenomena over time. Temporal conspicuousness map is computed for a video buffer and presents amount of motion in a frame. To compute such a map on the fly, we apply principal component analysis (PCA) procedure to discard redundant data and reconstruct the feature buffer extracted from video sequence.

Let us assume that we have a feature sequence $F = \{f_t, f_{t+1}, \dots, f_{t+n}\}$, where f_i is the feature vector of video frame at time i in column representation. In our

[1] S is a matrix of size 24×24.

(a) (b) (c)

Fig. 3. a) μ_f computed for 5 intensity sequences, b) first mean normalized frame $\tilde{f}_1$, whiter regions represent areas with movements (i.e. change over time), c) temporal conspicuousness map for the given feature sequence

implementation, a feature vector can contain frame intensity values or spatial conspicuousness map of a video frame. Initially, we subtract the mean value μ_f of each row of feature sequence; assuming it represents the static information of feature sequence. Consequently, $\tilde{F} = F - \mu_f$ will provide an approximation of movements (i.e. change over time) in the scene. Figure 3 depicts μ_f and an example of mean normalized frame for intensity features.

Afterwards, we apply eigendecomposition to covariance matrix of $\tilde{F}$, $\Sigma = EDE^T$, where E is the eigenvector and D is the diagonal eigenvalue matrices. Feature sequence consists of whitened back projected dimensionally reduced movement approximations (i.e. $\tilde{F}$) which is computed as follows:

$$\tilde{F}_p^w = \tilde{F} E_p D_p^{-1/2} E_p^T. \tag{4}$$

where E_p is the first p eigenvectors of covariance matrix and D_p is the diagonal matrix of the first p eigenvalues. This provides us orthonormal features and decreases redundancy which boosts salient regions. Finally, we define the temporal conspicuousness map as the average changes in the temporal window and compute it using

$$C_t(F) = \sum_i |\tilde{f}_{pi}^w|. \tag{5}$$

where $\tilde{f}_{pi}^w$ is the i_{th} column of $\tilde{F}_p^w$. Figure 3(c) visualizes a temporal conspicuousness map for a given sequence of intensity features.

2.3 Spatio-temporal Saliency Measure

The spatio-temporal saliency map is defined as the combination of temporal and spatial conspicuousness maps. We compute temporal stimulus change over spatial maps to obtain spatial-temporal maps and combine it with pure temporal information to gain a unique map. To this end, given a video buffer $V = \{v_t, v_{t+1}, \ldots, v_{t+n}\}$, where $v_i, \quad i \in [t, t+n]$ is the video frame at time i; we define the spatio-temporal saliency measure as follows:

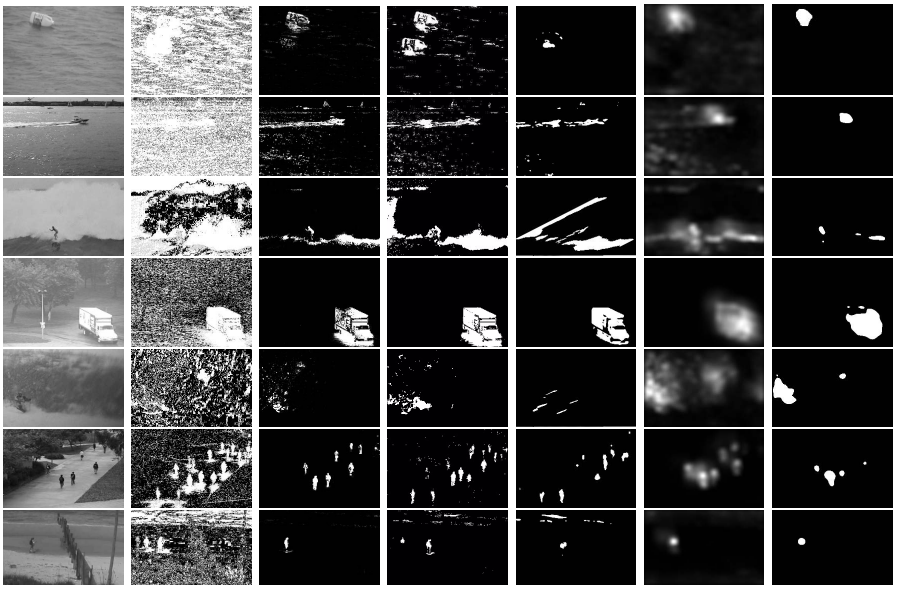


Fig. 4. Comparing background subtraction methods on the selected sequences; from left to right original frame, KDE, Adaptive GMM, ViBe, SC-SOBS and Proposed (NMF) method saliency map and segmented object by threshold value of 0.5. As depicted, KDE is vulnerable to high false positive noise when learning on a limited number of frames. Adaptive GMM and ViBe can handle videos with small number of learning frames with less false positive. The proposed method is the best of all in dealing with all the above sequences.

$$S(V) = w_0.C_t(V) + \sum_i w_i.C_t(C_s^i(V)), \tag{6}$$

where $C_s^i(V) = \{C_s(v_t), C_s(v_{t+1}), \ldots, C_s(v_{t+n})\}, \quad i = 1 \ldots m$ represents a buffer of spatial conspicuousness map obtained from i_{th} NMF and $\sum_{i=0}^{m} w_i = 1$ which are selected arbitrarily.

In order to compute final saliency map, we post process $S(V)$ by initially normalizing it to $[0, 1]$. We dilated the $S(V)$ to prevent severe attenuation on borders of salient region and applied a Gaussian filter to the attenuated map to have smooth saliency map. The procedure is summarized as follows:

$$SaliencyMap = G_{X,\sigma}((S(V) \oplus Disk_3)^{\alpha}), \tag{7}$$

where $\oplus$ is dilation operator, $Disk_3$ is a disk structure of element size 3, G is a Gaussian filter with $\sigma = 5$ and $\alpha = 10$ is the attenuation factor. To segment a foreground object, we apply a simple threshold where pixels with saliency greater than a threshold are labeled as foreground. Figure 4 depicts some saliency maps and segmentation result at threshold value of 0.5.

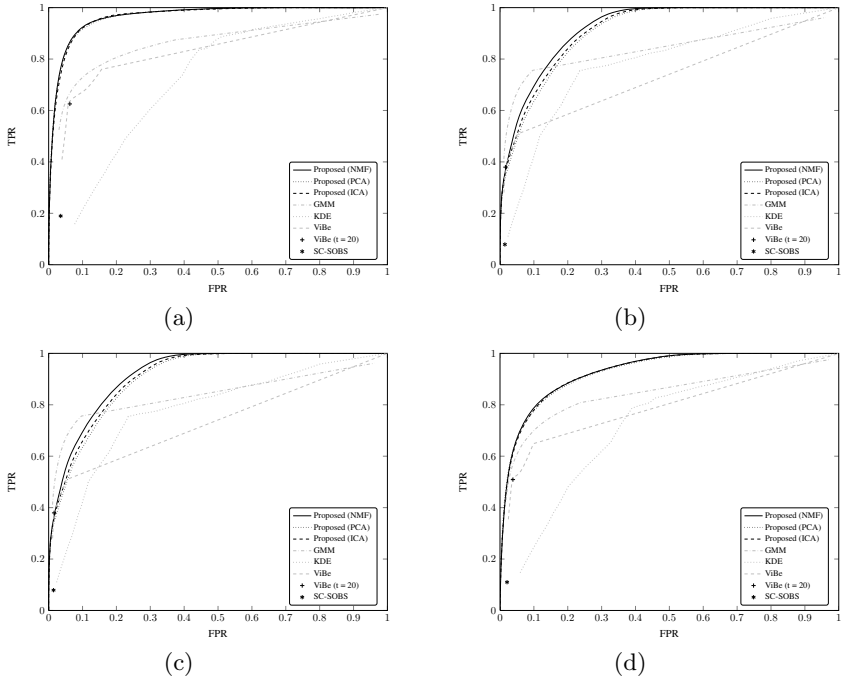


Fig. 5. ROC curve analysis for the selected sequences; a) sequences with dynamic background, b) sequences with environmental clutter (e.g. rain and snow), c) simple sequences, d) average performance of all the sequences

3 Experiments

In this section, we analyze the performance of the proposed method. To assess performance of NMF against other sparsification techniques, we compute the spatial conspicuousness maps using ICA and PCA as well as NMF. Afterwards, we compare them with background subtraction techniques; adaptive Gaussian Mixture Method of [17] (GMM), Elgammal et al. [18] none-parametric background modelling (KDE) and the recent state-of-the-art ViBe [30]and SC-SOBS [31]. Finally, we compare with saliency methods of [19,21].

We use UCSD background subtraction dataset to evaluate the performance. The data set is in gray-scale and there exists no empty background sequences which are required for learning a background model. Hence, we initialise the background subtraction algorithms with the first five frames of the sequence.

We compare the methods in terms of Receiver Operating Characteristic (ROC) curve which is obtained by computing false positive and true positive rates at different threshold values. It can be summarized in terms of Equal Error Rate (EER) which is the point false positive rate and true positive rate are equal.

Table 1. Comparing proposed method, Mahadevan[19] and Itti[21] for selected sequences in Terms of EER and running time (T) for a frame of 240×320

	Sequence	**Mahadevan**[19]	**Itti**[21]	**Proposed Method (NMF)**
EER	Bottle	2%	5%	9%
	Boat	9%	9%	8%
	Surf	4%	30%	8%
	Rain	3%	10%	10%
	Ski	3%	26%	13%
	Pedestrian	7%	37%	13%
	Ocean	11%	42%	20%
Average EER		5.5%	22.7%	11.5%
Run time		37(sec)	–	0.07(sec)

Initially, we carried out a detailed evaluation of background subtraction techniques on three selected group of sequences. The first group consists of sequences with water in the background where water fluctuation and rapid change of background provides a dynamic background. Figure 5(a) compares the three methods using ROC curve analysis; as shown the proposed method has the best performance.

The second sequence group covers cluttered scenes where the background is static and the foreground is moving, but there exists some clutter similar to rain or snow. Figure 5(b) summarizes the performance using ROC curve analysis. Summarizing the ROC curves in terms of area under the curve (AUC), proposed method outperforms GMM by 0.09 having AUC of 0.9248 vs. 0.8346.

The last sequence set does not contain much clutter and background is static. Results are summarized in Figure 5(c). These experiments show that the proposed method outperforms all the background subtraction techniques; Figure 5(d) depicts the average performance of all the sequences which repeats the same conclusion.

As the aforementioned experiments show proposed method with NMF outperforms all the methods. Considering the background subtraction schemes, GMM performs better than the others. It has less false positive compared to KDE and ViBe. The same can be concluded from Figure 4 which depicts some frame examples of the above sequences. As can be seen, KDE suffers from strong noise and ViBe has evident shadow effect. It seems SC-SOBS has not been able to converge using limited number of training frames. All the methods are performing acceptably well on the rain sequence where the initial five frames contain no object of interest.

In the second part of experiments, we compare with two saliency detection techniques of Mahadevan et al. [19] and Itti et al. [21]. They are two methods specifically developed for video sequences and both are applied to the same problem. Since access to the method of [19] was impossible for the authors, we borrow the results from the paper of [19]. They summarize the ROC curve results in terms of EER.

Table 2. Comparing running time and average EER on all the sequences of USCD background subtraction dataset

Method	Average EER	Real-time
GMM[17]	29.7%	Yes
KDE[18]	33.1%	Yes
ViBe	29.9%	Yes
Itti[21]	26.2%	Yes
Mahadevan[19]	**7.6%**	No
Proposed (NMF)	**17.7%**	**Yes**
Proposed (ICA)	19.14%	Yes
Proposed (PCA)	18.84%	Yes

Table 1 summarizes the EER of proposed method and those of Mahadevan and Itti for the above sequences. As can be observed, method of [19] is two times better than the proposed method in terms of EER. However, it has one drawback; [19] requires 37 seconds to process a frame of size 240×320. On the other hand, proposed method requires only 0.07 second to process the same frame which provides frame rate of 14 fps.

Table 2 compares all the mentioned methods except SC-SOBS and their properties over the USCD background subtraction dataset. The SC-SOBS algorithm was left out because of difficulties in tuning the parameters and we used only original threshold values that is not enough to have a ROC curve. As experiments showed, performance of proposed method with NMF sparsification technique is better than PCA and ICA. Considering the background subtraction techniques, KDE fails to adapt and learn background model having limited number of training frames. ViBe learns the background model better than KDE, however, it produces a shadow effect. So, both methods produce false positive and fail to model the background properly. Although GMM has a fairly better performance in comparison to KDE and ViBe, it still produces lots of false positive in case of dynamic scenes and falls behind the proposed method. The proposed method performs much better than that of Itti [21] in terms of EER. The only method that exceeds the performance of proposed method is of [19], though it is incapable of running in real-time which can be an important property for a practical background subtraction algorithm in many applications (e.g. surveillance, tracking, etc).

4 Conclusion

In this article, we introduce a spatio-temporal technique for saliency detection in image sequences. The method computes spatial conspicuousness maps using filter banks learned using none-negative matrix factorization. The proposed method combines the spatial conspicuousness maps and intensity features in a temporal cue using principal component to derive one saliency map.

We showed that NMF performs better than PCA and ICA when used to compute a spatial conspicuousness map. The method was compared with several background subtraction techniques. The detailed comparison showed that it outperforms all of them in complicated scenarios as well as simple ones. The method was also compared with two saliency techniques; the only technique that has better performance than the proposed method is among them. However, it requires a lot of computation power. Hence, we can conclude that the proposed method has the best performance among those algorithms capable of running in real-time.

References

1. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
2. Rahtu, E., Kannala, J., Blaschko, M.B.: Learning a category independent object detection cascade. In: IEEE International Conference on Computer Vision (2011)
3. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II-37–II-44 (2004)
4. Kanan, C., Cottrell, G.: Robust classification of objects, faces, and flowers using natural image statistics. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2472–2479 (2010)
5. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient Region Detection and Segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
6. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 366–379. Springer, Heidelberg (2010)
7. Frintrop, S.: General object tracking with a component-based target descriptor. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 4531–4536 (2010)
8. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. Trans. Img. Proc. 19(1), 185–198 (2010)
9. Lu, T., Yuan, Z., Huang, Y., Wu, D., Yu, H.: Video retargeting with nonlinear spatial-temporal saliency fusion. In: Proceedings of the 2010 IEEE 17th International Conference on Image Processing (2010)
10. Jacobson, N., Lee, Y.L., Mahadevan, V., Vasconcelos, N., Nguyen, T.: A novel approach to fruc using discriminant saliency and frame segmentation. IEEE Transactions on Image Processing 19(11), 2924–2934 (2010)
11. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: IEEE 12th International Conference on Computer Vision, pp. 2232–2239 (2009)
12. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: 2010 IEEE Conference Computer Vision and Pattern Recognition (CVPR), pp. 2376–2383 (2010)

13. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection: A systematic survey. Recent Patents on Computer Science 4(3), 147–176 (2011)
14. Calderara, S., Melli, R., Prati, A., Cucchiara, R.: Reliable background suppression for complex scenes. In: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, pp. 211–214 (2006)
15. Heikkilä, J., Silvn, O.: A real-time system for monitoring of cyclists and pedestrians. Image and Vision Computing 22(7), 563–570 (2004)
16. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. xxiii+637+663 (1999)
17. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. 27(7), 773–780 (2006)
18. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
19. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 171–177 (2010)
20. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision 8(7) (2008)
21. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49(10), 1295–1306 (2009)
22. Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
23. Tsotsos, J.K., Bruce, N.D.B.: Saliency based on information maximization. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems, vol. 18, pp. 155–162. MIT Press (2006)
24. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision 8(7) (2008)
25. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
26. Hoyer, P.O.: Modeling receptive fields with non-negative sparse coding. Neurocomputing 52–54, 547–552 (2003)
27. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. 5, 1457–1469 (2004)
28. Rajapakse, M., Wyse, L.: Nmf vs ica for face recognition. In: Guo, M. (ed.) ISPA 2003. LNCS, vol. 2745, pp. 605–610. Springer, Heidelberg (2003)
29. Olmos, A., Kingdom, F.A.A.: A biologically inspired algorithm for the recovery of shading and reflectance images. Perception 33, 1463–1473 (2004)
30. Barnich, O., Van Droogenbroeck, M.: Vibe: A universal background subtraction algorithm for video sequences. IEEE Transactions on Image Processing 20(6), 1709–1724 (2011)
31. Maddalena, L., Petrosino, A.: The sobs algorithm: What are the limits. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 21–26 (2012)