Salient objects detection in dynamic scenes using color and texture features

Satya M. Muddamsetty, Désiré Sidibé, Alain Trémeau, Fabrice F Mériaudeau

▶ **To cite this version:**

Satya M. Muddamsetty, Désiré Sidibé, Alain Trémeau, Fabrice F Mériaudeau. Salient objects detection in dynamic scenes using color and texture features. Multimedia Tools and Applications, 2017, 29, pp.1-14. 10.1007/s11042-017-4462-y . hal-01497204

HAL Id: hal-01497204

https://u-bourgogne.hal.science/hal-01497204

Submitted on 25 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Salient Objects Detection in Dynamic Scenes Using Color and Texture Features

Satya M. Muddamsetty · Désiré Sidibé · Alain Trémeau · Fabrice Mériaudeau

Received: date / Accepted: date

Abstract Visual saliency is an important research topic in the field of computer vision due to its numerous possible applications. It helps to focus on regions of interest instead of processing the whole image or video data. Detecting visual saliency in still images has been widely addressed in literature with several formulations. However, visual saliency detection in videos has attracted little attention, and is a more challenging task due to additional temporal information. A common approach for obtaining a spatio-temporal saliency map is to combine a static saliency map and a dynamic saliency map. In our work, we model the dynamic textures in a dynamic scene with local binary patterns to compute the dynamic saliency map, and we use color features to compute the static saliency map. Both saliency maps are computed using a bio-inspired mechanism of human visual system with a discriminant formulation known as center surround saliency, and are fused in a proper way. The proposed model has been extensively evaluated with diverse publicly available datasets which contain several videos of dynamic scenes, and comparison with state-of-the art methods shows that it achieves competitive results.

Keywords Visual saliency · Dynamic textures · Salient objects detection · Local binary patterns

corresponding author: D. Sidibé: E-mail: dro-desire.sidibe@u-bourgogne.fr

S. M. Muddamsetty · D. Sidibé · F. Mériaudeau
Université de Bourgogne - France Comté; Le2i - UMR CNRS 6306

A. Trémeau
Unversité Jean Monnet, Laboratoire Hubert Curien UMR CNRS 5116, St-Etienne, France

1 Introduction

Visual attention is one of the useful concepts for humans in their daily life and it holds an important place in computer vision applications such as object detection [1], image segmentation [2], robotic navigation and localization [3], video surveillance [4], object tracking [5], image re-targeting [6] and image/video compression [7]. For example, consider a visual scene which contains many objects with various visual characteristics such as shape, color, size and texture. Some of the objects might be moving while others are static. Despite the huge amount of available information, the visual information reaching our eyes is limited as we cannot acquire the whole scene at a time. Thus we perceive only a small part of the visual field and the remaining part looks blurry to us. This smaller part of the visual field is perceived clearly with maximum acuity. The mechanism in the brain that determines which part of the multitude of sensory data is currently of most interest is called selective attention. It is basically a process to detect a scene's region which is different from the surroundings. Understanding this mechanism is an active research area in cognitive sciences.

Visual attention is generally processed in two approaches which are bottom-up approach and top-down approach. Bottom-up attention approach is stimulus driven and is derived solely from the conspicuousness of regions in a visual scene. Top-down attention approaches are goal driven and refer to voluntary allocation of attention to certain features, objects or regions in space [8]. Bottom-up approach is more thoroughly investigated than top-down attention approach because the data-driven stimuli are easier to control than cognitive factors such as knowledge and expectations [9].

While saliency detection is a widely studied problem, most of the existing techniques are limited to the analysis of static images. A recent survey of state-of-art methods can be found in [10,28] and these approaches cannot be simply extended to the analysis of videos sequences. Indeed, a video contains strong spatial-temporal correlation between the regions of consecutive frames. Furthermore, the motion of foreground objects dramatically changes the importance of the objects in a scene which leads to a different saliency map of the frame representing the scene. In addition, we know that natural scenes are composed of several dynamic entities such as moving trees, waves in water, fog, rain, snow and different illumination variations. Additional camera motion along with dynamic entities further complicates the detection of foreground objects. All these characteristics make video processing for saliency evaluation a challenging task. However, detecting salient regions and salient objects in complex dynamic scenes would be helpful in applications such as tracking, robotic navigation and localization and many more. A majority of the existing spatio-temporal saliency models [4,11,12] uses optical flow methods to process the motion information. In these methods, motion intensity of each pixel is computed and the final saliency map represents the pixels which are moving against the background. Optical flow based methods can work when the scene studied has simple background and fail with complex background scenes.

To overcome the challenges of natural dynamic scenes, we propose a new spatio-temporal saliency detection method in this paper. Our method is based on local binary patterns (LBP) for representing the scene as dynamic textures. The dynamic textures are modeled using local binary patterns in orthogonal planes (LBP-TOP) which is an extension of the LBP operator in temporal direction [13]. Our contributions are threefold. First, we apply a center-surround mechanism to the extracted dynamic textures in order to obtain a measure of saliency in different directions. Second, we propose to combine color and texture features. In our model, the spatial saliency map is computed using color features, and the temporal saliency map is computed using dynamic textures from LBP in two orthogonal planes. The different saliency maps are then fused to obtain a final spatio-temporal saliency map. Finally, we evaluate our spatio-temporal saliency detection method on two large and diverse datasets which, respectively contain salient objects and human eye fixations as a ground truth.

The rest of the paper is organized as follows. In Section 2, we review some of the spatio-temporal saliency detection methods presented in literature. In Section 3, we describe the proposed spatio-temporal saliency model based on LBPTOP and color features. Section 4, shows performance evaluation of our method and comparison with other approaches on two different datasets containing segmented salient objects and eye tracking data. Finally, Section 5 gives concluding remarks.

2 Related Work

In this section, we provide a brief description of some of the saliency models described in literature, which all follow the bottom-up approach principles. In [1], authors proposed an information theoretic spatio-temporal saliency model which is computed from spatio-temporal volumes. In this method the spatial and temporal saliency are calculated separately and they are fused with a dynamic fusion method. Marat *et al.* [11] proposed a space-time saliency algorithm which is inspired by the human visual system. First, a static saliency map is computed using color features, and a dynamic saliency map is computed using motion information derived from optical flow. The two maps are then fused to generate space-time saliency map. In a similar way, Tong *et al.* [4] proposed a saliency model which is used for video surveillance. The spatial map is computed based on low level features and the dynamic map is computed based on motion intensity, motion orientation and phase.

A phase spectrum approach is proposed by Guo and Zhang [7]. In this method, motion is computed by taking the difference between two frames, and is combined with color and intensity. The features are put together using a quaternion representation and Quaternion Fourier Transform (QFT) is applied to get final saliency map. Kim *et al.* [15] presented a salient region detection method for both images and videos based on center-surround hypothesis. They used edge and color orientations to compute the spatial saliency. The dynamic saliency is computed by taking the absolute difference between the center and surround temporal gradients and is finally fused with the spatial map. Zhou *et al.* [16] proposed a dynamic saliency model to detect moving objects against dynamic backgrounds. This algorithm is based on the fact that the displacement of the foreground and the background can be represented by the phase change of the Fourier spectra, and the motion of background objects can be extracted by phase discrepancy in an efficient way.

In [17], Seo and Milanfar proposed a space-time saliency detection method which is based on a bottom-up framework and uses local regression kernels from a video as local features which differs from conventional filter responses. Local regression kernels capture the underlying local structure of the image very well even in the presence of significant distortions. In [17],

authors use a non parametric kernel density estimation for such features, which results in a saliency map constructed from a local self-resemblance measure computed using cosine similarity which indicates likelihood of saliency. A similar method is developed in [18], where the video patches are modeled using dynamic textures and saliency is computed based on discriminant center-surround.

Mancas *et al.* [19] proposed a bottom-up saliency method based on global rarity quantification. The model is based on a multi-scale approach using features extracted from optical flow, the final saliency map gives the rarity of the statistics of a given video volume at several scales. The authors in [20] proposed a dynamic saliency visual attention model based on the rarity of features. They introduced the Incremental Coding Length (ICL) to measure the perspective entropy gain of each feature using sparse coding techniques to represent features. Zhang *et al.* [21] proposed a saliency detection method based on Bayesian framework. The authors suggest that the pre-attentive process must estimates the probability of a target given the visual features at every location in the visual field to achieve the goal for detecting potentially important targets. This methods is based on a Bayesian framework from which bottom-up saliency emerges naturally, using image statistics derived from a large collection of natural images. Fu *et al.* [22] extended graph based approaches for saliency detection in videos by combining static appearance and motion cues into the graph construction.

Most of these methods fail to address complex scenes. In particular, methods based on optical flow fail to compute accurate dynamic saliency maps for scenes with highly textured backgrounds as will be shown in the experimental results in Section 4.

3 Spatio-temporal saliency detection using texture and color features

This section descibes the proposed spatio-temporal saliency detection method for dynamic scenes using LBP for describing the dynamic textures (DT) and color features for computing the static saliency. We first describe a method using only LBP feature computed in three orthogonal planes, and then show that using color features in combination with texture features produce better saliency maps.

3.1 Spatio-temporal saliency detection using LBPTOP descriptor

Dynamic or temporal textures are textures with motion that exhibit some stationary properties in time. The major difference between a DT and an ordinary texture is that the notion of self-similarity, central to conventional image texture, is extended to the spatio-temporal domain, thus a DT combines appearance and motion simultaneously [23]. Dynamic textures encompass the different difficulties of dynamic scenes such as moving trees, snow, rain, fog, crowd etc. Therefore, we use DT to model the varying appearance of dynamic scenes with time.

Several approaches have been developed to represent dynamic textures and a review of these methods can be found in [23]. In our work, we model DT using local binary patterns computed in orthogonal planes (LBPTOP) [13]. The LBPTOP operator extends LBP to temporal domain by computing the co-occurrences of local binary patterns on three orthogonal planes such as XY, XT and YT. The XT and YT planes provide information about the space-time transitions and the XY plane provides spatial information. These three orthogonal planes intersect at the center pixel. LBPTOP considers the feature distributions from each separate plane and then concatenates them into a single histogram.

In this work, we compute spatio-temporal saliency using a center-surround (CS) mechanism. CS is a discriminant formulation in which the features distribution of the center of visual stimuli is compared with the feature distribution of surrounding stimuli.

For each pixel location $l = (x_c, y_c)$, we extract a center region r_C and a surrounding region r_S both centered at l. We then compute the feature distributions $\mathbf{h}_c$ and $\mathbf{h}_s$ of both regions as histograms and define the saliency of pixel l as the dissimilarity between these two distributions. More specifically, the saliency $S(l)$ of pixel at location l is given by:

$$S(l) = \chi^2(\mathbf{h}_c, \mathbf{h}_s) = \sum_{i=1}^{B} \frac{(\mathbf{h}_c(i) - \mathbf{h}_s(i))^2}{(\mathbf{h}_s(i) + \mathbf{h}_c)/2}, \quad (1)$$

where $\mathbf{h}_c$ and $\mathbf{h}_s$ are the histograms distributions of r_C and r_S respectively, B is the number of bins of the histogram, and χ^2 is the Chi-square distance measure.

Note that we separately apply center-surround mechanism to each of the three planes XY, XT and YT. Hence, we compute three different saliency maps based on the three distributions derived from LBPTOP.

The final step of our method consists in fusing the previous three maps into a single spatio-temporal saliency map. This is done in two steps. First, the two maps containing temporal information, i.e. the saliency maps

from XT and YT planes, are fused to get a dynamic saliency map. Then, this dynamic saliency map is fused with the static saliency map from the XY plane. As shown in [12], the fusion method affects the quality of the obtained final spatio-temporal saliency map.

It is worth mentioning that the fusion of both maps into a single spatio-temporal saliency map can be considered as a multiview information fusion problem for which several approaches have been proposed in literature [24,25]. The main idea of those techniques is to treat each feature as a different view or a different projection of the data, and make use of the consistency and redundancy of different views to achieve better performance. In [24] it is shown that mutliview learning methods are based on the two main principles, which are consensus and complementary principles. The first principle aims to maximize the agreement on distinct multiple views, while the second one states that each view contains some information that other views do not have. Many multiview learning methods have been developed in recent years and the interested reader is referred to [26,24] for an overview.

In this work, we adopt the simple Dynamic Weighted Fusion (DWF) method, which has shown best performance in a recent evaluation [12]. This fusion scheme produces a weighted combination of both maps and the weights are adapted to the characteristics of the dynamic scene. In DWF the weights are calculated by computing a ratio between the means of both the maps to combine, so they are updated from frame to frame. Let S_{XT} and S_{YT} be the saliency maps obtained from the XT and YT planes respectively. They are fused into a dynamic saliency map M_D as follows:

$$M_D = \alpha_D S_{YT} + (1 - \alpha_D) S_{XT}, \tag{2}$$

where $\alpha_D = \frac{mean(S_{YT})}{mean(S_{XT}) + mean(S_{YT})}$.

The obtained dynamic map M_D and the static map $M_S = S_{XY}$ are fused in a similar manner.

3.2 Spatio-temporal saliency detection using color and texture feature

Since the final spatio-temporal saliency map is obtained as a fusion of the static and dynamic saliency maps, a proper static saliency map is needed in order to get an accurate spatio-temporal saliency map. In the previous approach, the spatial saliency map derived from the XY plane fails to highlight salient objects of some scenes because LBPTOP does not use color features. Threfore, we replace the LBP features computed in the XY plane by color features, since color is one of the salient feature in visual attention. In particular, we compute the spatial saliency map based on color features using the context-aware method of Goferman *et al.* [27] since this saliency detection method was shown to achieve best performance in a recent evaluation [28].

3.2.1 Spatial saliency

In our work, we used a saliency detection method based on context information [27]. Our choice is motivated by the fact that this method proves to be the best in a recent evaluation of saliency detection methods [28]. We only give a brief description of the method here, and we refer the interested reader to [27] for more details.

The saliency is computed in three steps. In the first step, local and global single scale saliency is computed for each pixel i in an image. A pixel i is considered salient if the appearance of the patch p_i centered at pixel i is distinctive with respect to all other image patches. The dissimilarity measure between the patches p_i and p_j is defined by:

$$d(p_i, p_j) = \frac{d_{color}(p_i, p_j)}{1 + c.d_{position}(p_i, p_j)}, \tag{3}$$

where d_{color} represents the Euclidean distance between the vectorized patches p_i and p_j of sizes 7×7 in CIElab color space which are normalized to the range $[0, 1]$, and $d_{position}$ is the Euclidean distance between the position of patches p_i and p_j. c is a constant scalar value set to $c = 3$ in our experiments (changing the value of c does not significantly affect the final result).

To evaluate a patch's uniqueness, there is no need to incorporate its dissimilarity to all the image patches. So for every patch p_i, we search for the K most similar patches q_k, $k = 1, \ldots, K$, in the image. The pixel i is considered salient when its dissimilarity $d(p_i, q_k)$ is high $\forall k \in [1, K]$.

In the second step, a multi-scale saliency is computed by considering different scales of the processed image. These multiple scales are utilized by representing each pixel i by the set of multi-scale image patches centered at it. The pixel i is considered as salient if it is consistently different from other pixels in multiple scales.

The final step includes the immediate context of the salient object. The immediate context suggests that areas that are close to the foci of attention should be explored significantly more than far-away regions. The visual context is simulated by extracting the most attended localized areas at each scale.

3.2.2 Spatio-temporal saliency map

The temporal saliency is computed as mentioned in Section 3.1. However, we consider here only two planes

XT and YT which gives information only in the temporal direction. The LBP features are extracted in XT and YT planes and two saliency maps are computed in both planes separately. These two maps are fused into a single dynamic saliency maps using the DWT fusion scheme as in Eq. 2.

Finally, the obtained spatial and temporal saliency maps, respectively M_S and M_D, are fused into the final spatio-temporal saliency map as:

$$M_{ST} = \alpha M_D + (1 - \alpha) M_S, \tag{4}$$

with $\alpha = \frac{mean(M_D)}{mean(M_D)+mean(M_S)}$, and S_{ST} the final spatio-temporal saliency map.

The last step of our method consists in applying a post-processing scheme to suppress the isolated pixels or group of pixels with low saliency values. We start this post-processing by finding pixels whose saliency value is above a defined threshold (0.5 in our experiments, the final saliency map M_{ST} is normalized to have values in $[0, 1]$). Then, we compute the spatial distance $D(x, y)$ from each pixel to the nearest non-zero pixel in the thresholded map. The spatio-temporal saliency map M_{ST} is finally refined using the following equation:

$$M_{ST}(x, y) = e^{\frac{-D(x,y)}{\lambda}} \times M_{ST}(x, y), \tag{5}$$

where λ is a constant set to $\lambda = 0.5$. We study the influence of this last parameter in the experimental results section.

4 Experimental evaluations

In this section we describe the experiments conducted to evaluate the efficiency of the proposed model. We performed two experiments to test the performance of the method in locating interesting foreground objects in complex scenes, and on the task of predicting human observers fixations. Firstly, we use a publicly available dataset of dynamic scenes [18] which contains ground truth segmentation of the salient objects for each frame of a sequence, thus allowing us to evaluate the ability of the method in detecting foreground objects in a complex scene. Secondly, we evaluate our model on another dataset in which the ground truth is given as eye tracking data, i.e. human observers fixations. This evaluates the performance of the model in predicting human fixation when viewing a video. The performances of the proposed method are also compared with various state-of-the-art methods.

4.1 Evaluation datasets and metrics

To evaluate the different spatio-temporal saliency models, we have selected two publicly available complex video scenes datasets: SVCL dataset [18] and ASCMN dataset [29]. The SVCL dataset, contains natural videos which are composed of dynamic entities such as waving trees, crowd, moving water, waves, snow and smoke filled environments. This dataset contains manually segmented objects for each frame which served as ground truth data.

The second dataset, ASCMN [29] is a collection of videos from various sources and provides data which cover a wider spectrum of video types. It contains totally 24 videos, together with eye tracking data collected from 13 human observers using eye tracking apparatus. The dataset is divided into 5 classes of sequences: *abnormal*, *surveillance*, *crowd*, *moving* and *noise*.

We use two evaluation metrics which are Area Under ROC Curve (AUC) [30] and Kullback-Leibler Divergence (KL-DIV) [31]. While only one of these measures is used in most of the previous works, in our experimental evaluation we use both measures to ensure that the discussion about the results is as independent as possible from the choice of the metrics.

AUC is used for assessing the degree of similarity of two saliency maps, and KL-DIV is used to estimate whether the saliency map produced by a saliency model matches human fixations. AUC varies from zero to one, with higher value indicating good performance, while KL-DIV varies from zero to infinity with zeros value indicating that two probability density functions are strictly equal.

4.2 Experiment 1: detection of salient objects in dynamic scenes

In this section, we evaluate the performance of the proposed spatio-temporal saliency detection algorithm in decting salient objects in complex dynamic scenes. We used the SVCL dataset for this experiment and compare our proposed methods with other state-of-the-art techniques. We compare two versions of our method which are LBPTOP (the method using only texture features from LBPTOP operator) and LBPTOP-COLOR (the method combining color features and LBPTOP features), and three existing methods: a method using optical flow to compute motion features (OF) [12], the self-resemblance method (SR) [17] and the phased discrepancy based saliency detection method (PD) [16]. For the last three methods, we use codes provided by the authors. For LBPTOP based saliency, we use center-surround mechanism described in Section 3.1 with a center region of size 17×17 and a surround region of size 97×97, and we extract LBP features from a temporal volume of six frames.

We evaluate the different spatio-temporal saliency detection methods by generating Receiver Operating Characteristic (ROC) curves and evaluating the Area Under ROC Curve (AUC). For each method, the obtained spatio-temporal saliency map is first normalized to the range $[0, 1]$, and binarized using a varying threshold $t \in [0, 1]$. With the binarized maps, we compute the true positive rate and false positive rate with respect to the ground truth data.

The post-processing step described in Section 3.2.2 is important in order to obtain good final saliency maps. It basically lower the final saliency value of pixels far away from all pixels with saliency value above a defined threshold. The parameter λ in Eq. (5) controls the importance of the attenuation. In this experiment, we set the value $\lambda = 0.5$ as it is, in average, the best value for all tested sequences.

The results obtained with all sequences by the different saliency detection methods are summarized in Table 1. As can be seen in Table 1, the proposed method combining color and texture features (LBPTOP-COLOR) achieves the best overall performance with an average AUC value of 0.914 for all twelve sequences. The optical flow based method (OF) achieves an average AUC value of 0.907, whereas as self-resemblance (SR), phase discrepancy (PD) and the method using texture features only(LBOTOP) achieve lower average AUC values, respectively 0.843, 0.837 and 0.745. These results confirm the observation that the combination of color features with LBP features produces better saliency map. In fact, the proposed method fusing color and LBP features gives an average AUC value which is 22% higher that the value with LBPTOP features alone.

When we analize the individual sequences, we see that the best and least performances are obtained with the *Boats* and *Freeway* sequences, respectively, with average AUC values of 0.9394 and 0.7398 for all five saliency detection methods. The *Boats* sequence shows good color and motion contrasts, so both static and dynamic maps are estimated correctly, and all spatio-temporal saliency detection methods perform well. Note however that the texture only based method (LBP-TOP) gives slightly lower accuracy than other techniques. On the other hand, the color contrast of the *Freeway* sequence is very limited. So getting a correct static saliency map is difficult with this sequence whereas the quality of the final spatio-temporal saliency map relies on the dynamic map. The best performing method with this sequence is the LBPTOP based technique with an average AUC value of 0.868, while optical-flow based technique achieves an average AUC value of only 0.545. This example illustrates that using LBP features to represent dynamic textures, and to compute the dynamic saliency map, gives very good results. The ROC curves comparing performances of the different methods on these two sequences are shown in Fig. 1 and Fig. 2.

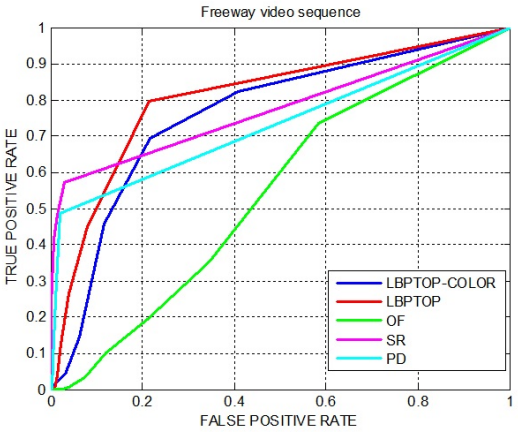


Fig. 1 Quantitative comparision with *Freeway* sequence from SVCL dataset and AUC metric.

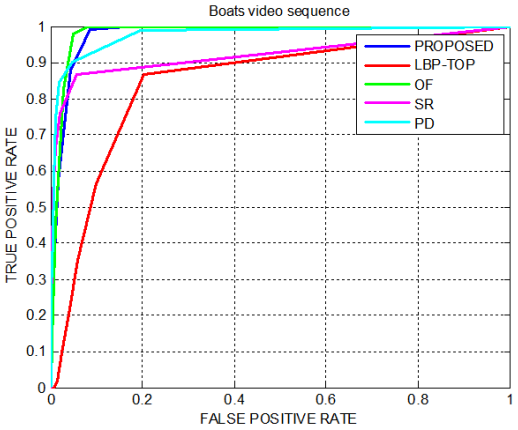


Fig. 2 Quantitative comparision with *Boats* sequence from SVCL dataset and AUC metric.

4.3 Experiment 2: prediction of human fixations

In this section we evaluate the performance of the proposed method in predicting human fixations using the ASCMN dataset [29] which contains 24 videos divided into five classes. We compare our proposed spatio-temporal saliency detection methods, LBPTOP and LBPTOP-COLOR, with four state-of-the-art methods which are the incremental coding length (ICL) method [20], the method based on natural images statistics (SUN) [32], the self-resemblance method (SR) [17], and the method of Mancas et al. [19].

For this second experiment, the parameter λ in Eq. (5) is set to $\lambda = 0.2$ for the proposed LBPTOP-COLOR method as it is the best value for all tested sequences.

Sequence	LBPTOP-COLOR	LBPTOP	OF [12]	SR [17]	PD [16]	Avg AUC
Birds	0.9586	0.7680	0.9664	0.9379	0.8221	0.8906
Boats	0.9794	0.8358	0.9827	0.9227	0.9765	**0.9394**
Bottle	0.9953	0.9413	0.8787	0.9961	0.8285	0.9279
Cyclists	0.9317	0.6737	0.9602	0.8682	0.9551	0.8777
Chopper	0.9717	0.9427	0.9850	0.7447	0.6470	0.8582
Freeway	0.7775	0.8684	0.5456	0.7760	0.7318	**0.7398**
Peds	0.9552	0.7376	0.9512	0.8603	0.8548	0.8718
Ocean	0.9271	0.8513	0.7810	0.8016	0.8235	0.8369
Surfers	0.9674	0.7489	0.9545	0.9455	0.9352	0.9103
Skiing	0.8389	0.3787	0.9796	0.8872	0.9367	0.8042
Jump	0.8957	0.6960	0.9481	0.8321	0.6616	0.8067
Traffic	0.7693	0.6088	0.9615	0.5491	0.8720	0.7521
Avg AUC	**0.9140**	**0.7453**	**0.9079**	**0.8434**	**0.8371**	

Table 1 Evaluation of spatio-temporal saliency detection methods using the SVCL dataset. LBPTO-CPOLOR (proposed method with color and LBP features), LBPTOP (propose method with LBP features only), OF (Optical Flow based), SR (Self-Resemblance) and PD (Phase Discrepancy).

We compare the different saliency detection methods both in terms of the evaluation metric used and the type of the video sequence used.

Table 2 summarizes the results obtained by the different saliency detection methods for all the twenty four video sequences of the dataset, using AUC and KL-DIV metrics respectively. First of all, we can see that the relative performances of the different methods depends on the evaluation metric used. This justify our idea of using more than one metric to ensure that the discussion about the results is as independent as possible from the choice of the metrics.

In terms on evaluation metrics, for AUC the higher the value the better is the performance of a method. On the contrary, for the KL-DIV measure, the lower the value the better the performance of a method. Table 2 shows that the proposed method combining color and texture features (LBPTOP-COLOR) achieves an average AUC value of 0.64, which is higher that the performance of ICL, LBPTOP and SUN methods which achieve average AUC values of 0.63, 0.53 and 0.61 respectively. However, LBPTOP-COLOR has a lower performance compared to MANCAS and SR methods which achieve average AUC value of 0.68 and 0.66 respectively. When using KL-DIV metric, the distributions given by the eye fixations points and the saliency maps produced by the model are first and the KL-divergence measure is computed between these two distributions to estimate whether the saliency map produced by a saliency model matches human fixations. From Table 2, we can see that LBPTOP-COLOR method achieves the second best result, being outperformed only by SR. However, we can also see that all saliency methods gives comparable results in terms of KL-DIV measure. A visual comparison of the results obtained with different methods is shown in Fig. 3.

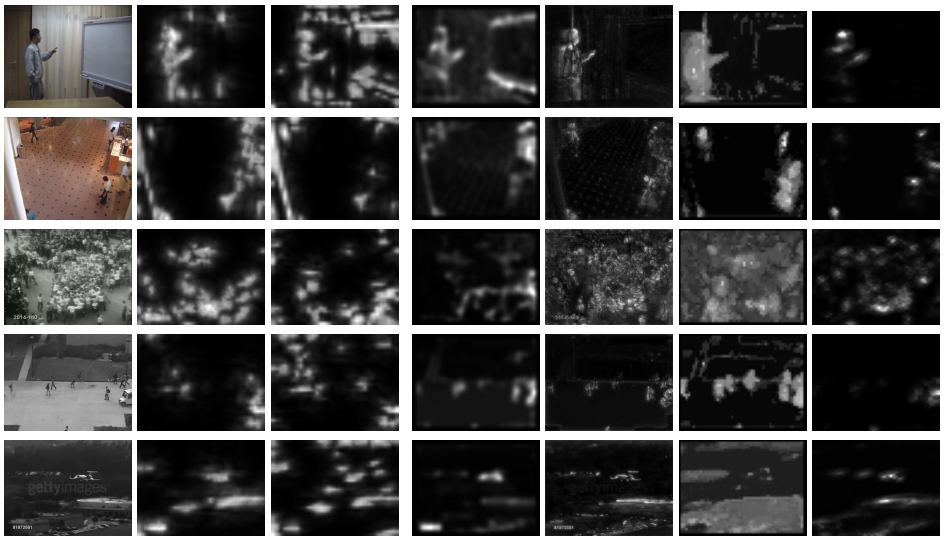
MODELS	mean AUC	mean KL
LBPTOP-COLOR	64%	1.5860
LBPTOP	53%	1.6059
ICL [20]	63%	1.5899
SUN [32]	61%	1.587
MANCAS [19]	68%	1.6158
SR [17]	66%	1.5662

Table 2 Evaluation of saliency detection methods using the ASCMN dataset with two evaluation metrics

5 Conclusion

This paper describes a spatio-temporal saliency detection method in dynamic scenes based on the combination of color and texture features. Color features are used to compute a static saliency map for each frame of a sequence, and local binary patterns describing dynamic textures are used to find a dynamic map. The obtained two saliency maps are then fused into a spatio-temporal saliency map which can be used for objects segmentation. Extensive experiments with two large and diverse datasets show that the proposed method combining color and texture features performs significantly better than a method using LBP feature only, and also better than method based on optical flow estimation for the dynamic saliency computation. The proposed method can, in particular, deal with dynamic scenes with difficult background textures, but achieves lower results when the contrast is poor.

A possible extension of this work could be the integration of depth cues into the spatio-temporal saliency model. The current availability of RGB-D sensor makes this possible and we will investigate this in the future. Also, we could consider the fusion of static and dynamic saliency maps as a multiview information fusion problem and adopt a multiview learning approach.



(a) (b) (c) (d) (e) (f) (g)

Fig. 3 Visual comparison of spatio-temporal saliency detection of our methods and state of art methods on ASCMN dataset. (a) Original frame; (b) LBPTOP-COLOR; (c) LBPTOP; (d) ICL [20]; (e) SUN [32]; (f) MANCAS [19] and (g) SR [17]

References

1. L. Chang, P. C. Yuen, G. Qiu, Object motion detection using information theoretic spatio-temporal saliency, Pattern Recogn. 2009 42 (11) 2897–2906.
2. R. Achanta, F. Estrada, S. Susstrunk, S. Hemami, Frequency-tuned salient region detection, Computer Vision and Pattern Recognition, 2009 1597–1604.
3. C. Siagian, L. Itti, Biologically inspired mobile robot vision localization, IEEE Transactions on Robotics 25 (4) (2009) 861–873.
4. T. Yubing, F. A. Cheikh, F. F. E. Guraya, H. . Konik, A. Trmeau, A spatiotemporal saliency model for video surveillance, Cognitive Compuation,2011 Volume 3, Issue 1 241–263.
5. D. Sidibé, D. Fofi, F. Mériaudeau, Using visual saliency for object tracking with particle filters, in: EUSIPCO, 2010.
6. T. Lu, Z. Yuan, Y. Huang, D. Wu, H. Yu, Video retargeting with nonlinear spatial-temporal saliency fusion, in: ICIP,2010.
7. C. L. Guo, L. M. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE TIP 19 (1) (2010) 185 –198.
8. Y. Pinto, A. R. van der Leij, I. G. Sligte, V. A. F. Lamme, H. S. Scholte, Bottom-up and top-down attention are independent, Journal of Vision 13 (3) (2013) 16.
9. S. Frintrop, Computational Visual Attention, Springer, 2011.
10. A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans on Pattern Analysis and Machine Intelligence, 35 (1) (2013) 185–207.
11. S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, IJCV,2009 82 (3) 231–243.
12. S. M. Muddamsetty, D. Sidibé, A. Trémeau, F. Mériaudeau, A performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes, in: ICIP, 2013.
13. G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 915–928.
14. L. ltti, C. Koch, E. Neibur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans on Pattern Analysis and Machine Intelligence,1998 20 1254–1259.
15. W. Kim, C. Jung, C. Kim, Spatiotemporal saliency detection and its applications in static and dynamic scenes, IEEE Trans. Circuits Syst. Video Techn,2011 21 (4) 446–456.
16. B. Zhou, X. Hou, L. Zhang, A phase discrepancy analysis of object motion, in: InProceeding of the 10th Asian Confernce of Computer Vision, 2011, pp. 225–238.
17. H. J. J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, Journal of vision 9 (12).
18. V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (1) (2010) 171 –177.
19. M. Mancas, N. Riche, J. Leroy, B. Gosselin, Abnormal motion selection in crowds using bottom-up saliency, in: IEEE ICIP 2011, pp. 229–232.
20. X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments., in: NIPS, Vol. 5, 2008, p. 7.
21. L. Zhang, M. H. Tong, T. K. Marks, H. Shan, G. W. Cottrell, Sun: A bayesian framework for saliency using natural statistics, Journal of vision 8 (7) (2008) 32.
22. K. Fu, I. Y. H. Gu, Y. Yun, C. Gong, J. Yang, Graph construction for salient object detection in videos, in: ICPR, 2014, pp. 2371–2376.
23. D. Chetverikov, R. Péteri, A brief survey of dynamic texture description and recognition, in: Computer Recognition Systems, Springer, 2005, pp. 17–26.
24. C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint (2013), pp. 1304.5634.
25. C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Trans. on PAMI 37 (12) (2015), pp. 2531–2544.
26. S. Sun, A survey of multi-view machine learning, Neural Computing and Applications, 23 (7) (2013), pp. 2013–2038.

27. S. Goferman, L. Zelnik-manor, A. Tal, Context-aware saliency detection, in: IEEE CVPR,2010.
28. A. Borji, M. M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. on Image Processing, 24 (12) (2015), pp. 5706–5722.
29. N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, T. Dutoit, Dynamic saliency models and human attention: A comparative study on videos, in: Computer Vision–ACCV 2012, Springer, 2013, pp. 586–598.
30. T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (8) (2006) 861–874.
31. O. Le Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior research methods 45 (1) (2013) 251–266.
32. L. Zhang, M. H. Tong, G. W. Cottrell, Sunday: Saliency using natural statistics for dynamic analysis of scenes, In-Proceedings of the 31st Annual Cognitive Science Conference (2009) 2944–2949.