Contrast Based Hierarchical Spatial-Temporal Saliency for Video

Trung-Nghia Le[1(✉)] and Akihiro Sugimoto[2]

[1] Department of Informatics, SOKENDAI (Graduate University for Advanced Studies), 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
ltnghia@nii.ac.jp
[2] National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
sugimoto@nii.ac.jp

Abstract. Predicting human attention for video is requires exploiting temporal knowledge included in the video. We propose a novel hierarchical spatial-temporal saliency model for video based on the center-surround framework using both static features and temporal features. Saliency cues are analyzed through a hierarchical segmentation model, and fused across multiple levels, yielding the spatial-temporal saliency map. An adaptive temporal window using motion information is also developed to combine saliency values of consecutive frames in order to keep temporal consistency across frames. Performance evaluation on several popular benchmark datasets validates that our method outperforms existing state-of-the-arts.

1 Introduction

Predicting human attention plays a significant part in computer vision, mobile robotics, and cognitive systems [1]. Saliency detection aims to simulate human attention by focusing on the most informative and interesting regions in a scene. Several computational models are developed for human gaze fixation prediction [2,3], which is important for understanding human attention; while others are proposed for salient object detection [4,5], which is useful for high-level vision tasks. In this work, although we focus on the salient object detection aspect, our proposed saliency model also achieves high performance in predicting human gaze fixation.

In the real world, visual information is usually composed of dynamic entities caused by egocentric movements or dynamics of the world. Particularly, in a dynamic scene, background always changes; different parts corresponding to different elements or objects can move in different directions with different speed independently. Therefore, predicting human attention for video is challenging because we have to incorporate a relationship of dynamics between consecutive frames. Attention models should have ability to fuse current static information and accumulated knowledge on dynamics from the past to deal with the dynamic nature of scenes including two properties: dynamic background and

T. Bräunl et al. (Eds.): PSIVT 2015, LNCS 9431, pp. 734–748, 2016.
DOI: 10.1007/978-3-319-29451-3_58

entities' independent motion. Several spatial-temporal saliency methods based on motion analysis are proposed for video [4,6]. Some of them can capture scene regions that are important in a spatial-temporal manner [4,7]. However, most of existing methods do not fully exploit the nature of dynamics in a scene. Temporal features expressing motion dynamics of objects in a scene between consecutive frames are not utilized in saliency detection process, either.

In order to effectively use knowledge on dynamics of background and objects in a video, we propose a flexible framework where pixel-based features and region-based features are fused to create a saliency detection method (c.f. Table 1). In this framework, static features and temporal features computed from pixel-based and region-based features are combined together in order to utilize both low-level features of each frame and consistency between consecutive frames. We also present a novel metric for motion information by estimating the number of referenced frames for each single object to keep temporal consistency across frames. Our method overcomes the limitation of the existing method [4] which uses a fixed number of referenced frames and does not concern motion of objects within a scene.

Table 1. Feature classification

	pixel-based feature	region-based feature
static feature	- color - intensity - orientation	- location - objectness
temporal feature	- flow magnitude - flow orientation	- movement

In our method, firstly, we execute a hierarchical segmentation. Saliency map for each segmentation level is then calculated via combination of contrast information and regional characteristics among segmented regions at the same scale level. Our feature maps are combinations of pixel-based features and region-based features. Particularly, pixel-based features consist of low-level image features such as color, intensity, or orientation as well as temporal features such as flow magnitude or flow orientation; while region-based features include spatial features such as location of an object or foreground object as well as movement of an object (c.f. Table 1). An adaptive sliding window in the temporal domain is proposed to relate salient values of frame sequences by exploiting motion information of a segmented region in each frame. Each region in each frame has a different number of referenced frames depending on its motion distribution. Experimental results using two public standard datasets i.e., the Weizmann standard dataset [8] and the SFU dataset [9], show that our proposed method outperforms the state-of-the-arts. Examples of generated saliency maps using our method are shown in Fig. 1.

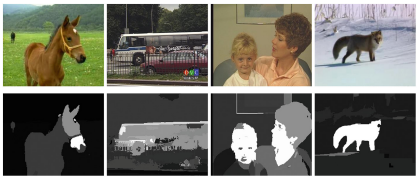


Fig. 1. Examples of our spatial-temporal saliency model. Top row images are original images. Bottom row images are the corresponding saliency maps using our method.

Our key contributions lie in twofold:

- The first one is that we show the framework, which integrates the contrast information together with regional properties. Although the proposed saliency model is developed based on a contrast based method presented by Zhou et al. [4], it significantly improves performance of the original work.
- The other is that we introduce a novel metric using motion information in order to keep temporal consistency between consecutive frames of each entity in a video. Our method also exploits the dynamic nature of the scene in term of independent motion of entities.

2 Related Work

Many computational models have been recently proposed for saliency detection. The majority of existing visual attention methods are developed using bottom-up computational algorithms, where low-level stimuli in scenes such as color, intensity, or edge are utilized in the center-surround contrast framework. For videos, several spatial-temporal saliency methods based on the center-surround framework are proposed. These methods measure the saliency of a pixel based on its contrast within a local context or the entire image. For instance, the framework proposed by Seo et al. [10] relies on center-surround differences between a local spatial-temporal cube and its neighboring cubes in space-time coordinates. In several center-surround schemes, motion between a pair of frames (e.g. optical flow), which is considered as a low-level feature channel, is used to compute local discrimination of the flow in a spatial neighborhood [4,11].

However, such contrast based saliency models may be ineffective when objects contain small-scale salient patterns; thus saliency could generally be misled by their complexity. Multi-level analysis and hierarchical models are developed to deal with salient small scale structure [4,5]. Some saliency models employ temporal coherence to principles of multi-scale processing to enhance performance [4].

Comparing with the previous work, our saliency method combines various features including primitive pixel-based features (color, intensity, orientation, and flow information) and region-based features (location, objectness and object's movement) through a hierarchical contrast calculation.

Saliency detection for video originates from applying an attention model to each frame of the video separately [12]. However, this kind of process does not achieve high effectiveness because temporal information across frames in the video is disregarded. The problem is even more challenging when dealing with dynamic scenes, where not only objects but also background always changes over time. Dynamics in a video is caused by different dynamic entities of natural scenes or by ego-motion of imaging sensors. Therefore, dynamic textures are integrated into discriminant center-surround saliency detection method to deal with scenes with highly dynamic backgrounds and moving cameras [13]. Accuracy for human egocentric visual attention prediction is also enhanced by adding information of camera's rotation, velocity and direction of movement into the bottom-up saliency model [6]. Differently from existing methods, our spatial-temporal saliency model uses motion information in order to keep temporal consistency across frames.

3 Hierarchical Spatial-Temporal Saliency Model

Figure 2 illustrates the process of our spatial-temporal saliency detection method. First of all, the streaming hierarchical method [14], which runs on arbitrarily long videos with constant, low memory consumption, is executed to hierarchically segment a video into spatial-temporal regions. In order to obtain regions at different scales, we initially construct a 5-level segmentation pyramid. Motion information as well as used features for each frame are extracted in each scale level. From these features, we build feature maps, including contrast information between regions and regional characteristics, in order to calculate saliency entities for regions in each scale level. After that, an Adaptive Temporal Window (ATW) is individually applied to each region to smooth saliency entities between frames by exploiting the motion information, yielding hierarchical saliency maps for each frame. Finally, a spatial-temporal saliency map is generated for each frame by fusing its hierarchical saliency maps.

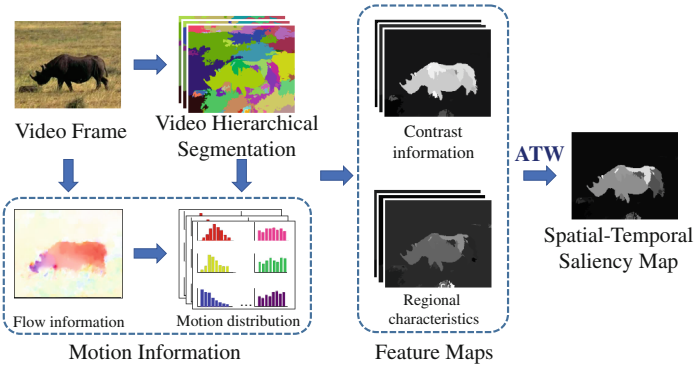


Fig. 2. Pipeline of the proposed spatial-temporal saliency method.

3.1 Saliency Entity Construction

Contrast Information. Human vision reacts to image regions with discriminative features such as unique color, high contrast, different orientation, or complex texture. To estimate attractiveness of regions in a video, contrast metric is usually used to evaluate sensitivity of elements in each frame. The contrast information is based on pixel-based features including static information such as color, intensity, or texture, and dynamic information such as magnitude or orientation of motion. A region with a high level of contrast against surrounding regions can attract human attention and is perceptually more important.

For the i-th region at the l-th scale of the segmentation pyramid at a frame, denoted by $r_{i,l}$, we compute its normalized color histogram in CIE Lab color space, denoted by $\chi_{i,l}^{pf_{col}}$, and distribution of lightness $\chi_{i,l}^{pf_{lig}}$. Gabor filter [15] is used to calculate orientation representation statistics $\chi_{i,l}^{pf_{ori}}$ of the region $r_{i,l}$.

Since human visual system is more sensitive to moving objects than still objects, temporal features are also compared between regions at the same segmentation level. Pixel-wise optical flow [16] is used to analyze motion between consecutive frames. Motion distribution of region $r_{i,l}$ is encoded in two descriptors: $\chi_{i,l}^{pf_{fmag}}$ is a normalized distribution of flow magnitude and $\chi_{i,l}^{pf_{fori}}$ is a normalized histogram of flow orientation.

The contrast of each region is measured as the sum of its feature distances to other regions at the same scale level in the segmentation pyramid with different weight factors:

$$S_{CI_{i,l}} = \sum_{pf} w_{pf} \sum_{j \neq i} |r_{j,l}| \, \omega\left(r_{i,l}, r_{j,l}\right) \left\| \chi_{i,l}^{pf} - \chi_{j,l}^{pf} \right\|, \tag{1}$$

where $|r_{j,l}|$ denotes the number of pixels in region $r_{j,l}$. $\left\| \chi_{i,l}^{pf} - \chi_{j,l}^{pf} \right\|$ is the Chi-Square distance [17] between two histograms, $pf \in \{pf_{col}, pf_{lig}, pf_{ori}, pf_{fmag}, pf_{fori}\}$ denotes one of the five features with corresponding weight w_{pf}. In this work, we use the same weights for all features. Regions with more pixels contribute higher contrast weight factors than those containing smaller number of pixels. $\omega\left(r_{i,l}, r_{j,l}\right)$ controls spatial distance influence between two regions $r_{i,l}$ and $r_{j,l}$:

$$\omega\left(r_{i,l}, r_{j,l}\right) = e^{\frac{-D\left(r_{i,l}, r_{j,l}\right)^2}{\sigma^2}},$$

where $D(r_{i,l}, r_{j,l})$ is the Euclidean distance between region centers and parameter σ controls how large the neighbors are. Finally, $S_{CI_{i,l}}$ is normalized to range $[0, 1]$.

Regional Characteristics. In addition to the contrast between regions, we also compute characteristics of each region based on region-based features. Human vision is biased toward specific spatial information of video such as center of the

frame or foreground objects, as well as movements of objects over time. Therefore, our region-based features are based on location, objectness, and movement metrics.

Human eye-tracking studies show that human attention favors the center of natural scenes when watching videos [18]. So, pixels close to the screen center could be salient in many cases. Our location feature is defined as:

$$\chi_{i,l}{}^{rf_{loc}} = \frac{1}{|r_{i,l}|} \sum_{j \in r_{i,l}} e^{\frac{-D(p_j,\bar{p})^2}{\sigma^2}},$$

where $|r_{i,l}|$ denotes the number of pixels in region $r_{i,l}$ and $D(p_j, \bar{p})$ is the Euclidean distance from each pixel p_j in the region to the image center $\bar{p}$.

The second characteristic is objectness of regions, which is based on differences of spatial layout of image regions [19]. Object regions are much less connected to image boundaries than background ones. In contrast, a region corresponding to background tends to be heavily connected to image boundary. In order to compute objectness of each region, each segmented image is first built as an undirected weighted graph by connecting all adjacent regions and assigning their weights as the Euclidean distance between their average colors in the CIE-Lab color space. The objectness feature of region $r_{i,l}$ is written as:

$$\chi_{i,l}{}^{rf_{obj}} = \exp\left(-\frac{BndCon^2(r_{i,l})}{2\sigma^2_{BndCo}}\right),$$

where $BndCon(r_{i,l})$ is the boundary connectivity of region $r_{i,l}$, which is calculated as the ratio of the length along the boundary of region $r_{i,l}$ to the square root of its spanning area. The length along the boundary of region $r_{i,l}$ is the sum of the Geodesic distance [19] from it to regions on the image boundary whereas its spanning area is the sum of the Geodesic distances from it to all regions in the image. σ_{BndCo} is a parameter and we set $\sigma_{BndCo} = 1$ like [19].

Moreover, to encode movement of objects, we capture any sudden speed change in motion of regions. Movement of a region is calculated as displacement of its spatial center over time:

$$\chi_{i,l}^{rf_{mov}} = e^{\lambda \Delta(r_{i,l})},$$

where $\Delta(r_{i,l})$ is the Euclidean distance between the centers of region $r_{i,l}$ in two consecutive frames.

The characteristics of the region $r_{i,l}$ are computed as the sum of its attribute values with different weight factors:

$$S_{RC_{i,l}} = \sum_{rf} w_{rf} \chi_{i,l}^{rf}, \tag{2}$$

where $rf \in \{rf_{loc}, rf_{obj}, rf_{mov}\}$ denotes one of the three features, with corresponding weight factor w_{rf}. In this work, we use the same weights for all features. Finally, $S_{RC_{i,l}}$ is normalized to range $[0, 1]$.

Feature Map Combination. Combining the contrast information and the regional characteristics, we obtain initial saliency entities for all segmentation levels separately:

$$S_{i,l} = S_{CI_{i,l}} S_{RC_{i,l}}. \tag{3}$$

Finally, the saliency entities are linearly normalized to fixed range $[0, 1]$ in order to guarantee that pixel with value 1 is the maxima of saliency.

3.2 Exploiting Motion Information

Differently from still images, video scenes include both spatial information and temporal information, which is considered as motion information of objects in the scenes. In addition, motion information plays an important role for human perception. Therefore, it is necessary to include motion information into the saliency map for video. To calculate motion information, we first use the pixel-wise optical flow proposed by C. Liu [16] to compute motion magnitude of each pixel in a frame, and then exploit distribution of motion magnitude in each region (c.f. Fig. 3).

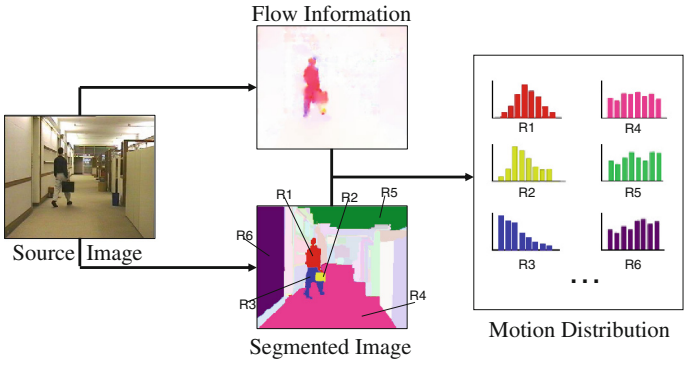


Fig. 3. Motion information calculation.

In a video, it is sometimes hard to distinguish objects from background because every pixel value always changes over time regardless that it belongs to an object or background. Moreover, motion analysis shows that different parts of objects move with various speed and, furthermore, background motion also changes with different speed and direction (c.f. flow information in Fig. 3). This causes fluctuation of object appearances between frames. To reduce this negative effect, saliency entities at each segmentation level at the current frame is combined with neighboring frames, resulting in smoothing saliency values over time. We propose to adaptively use a sliding window in the temporal domain for each region at each frame to capture speed variation by exploiting motion information

in the region. After this operation, salient values on contiguous frames become similar, and this generates robust temporal saliency:

$$\tilde{S}^t_{i,l} = \frac{1}{\Psi} \sum_{t'=t-\Phi^t_{i,l}}^{t} e^{\frac{-D(t,t')^2}{2{\Phi^t_{i,l}}^2\sigma^2}} S^t_{i,l}, \tag{4}$$

where $S^t_{i,l}$ measures saliency entity of region $r_{i,l}$ at frame t, $D(t,t')$ denotes the time difference between two frames, parameter σ controls how large the region at previous frames is. Ψ is the normalization factor of saliency value:

$$\Psi = \sum_{t'=t-\Phi^t_{i,l}}^{t} e^{\frac{-D(t,t')^2}{2{\Phi^t_{i,l}}^2\sigma^2}},$$

where $\Phi^t_{i,l}$ controls the number of participating frames in the operation, expressed as:

$$\Phi^t_{i,l} = Me^{-\mu^t_{i,l}\frac{\lambda}{\beta^t_{i,l}}}, \tag{5}$$

where M and λ are parameters. $\beta^t_{i,l} = \frac{\sigma^t_{i,l}}{\mu^t_{i,l}}$ is the coefficient variation measuring dispersion of motion distribution of each region. $\mu^t_{i,l}$ and $\sigma^t_{i,l}$ are the mean value and the standard deviation of the motion distribution of region $r_{i,l}$ at frame t.

3.3 Spatial-Temporal Saliency Generation

Normalized hierarchical saliency maps of different scales are combined to create a spatial-temporal saliency map SM by calculating the average over all hierarchical levels:

$$SM^t_p = \frac{1}{L}\sum_{l=1}^{L} \tilde{S}^t_{\Omega_l(p),l}, \tag{6}$$

where Ω_l is a function that converts pixel p to the region at scale level l where it belongs. Therefore, all operations are processed pixel-wisely. $\tilde{S}^t_{\Omega_l(p),l}$ measures hierarchical saliency value of each pixel generated in the l-th scale of the segmentation pyramid at frame t.

4 Experimental Setup

4.1 Dataset

We used Weizmann human action database [8] and SFU eye-tracking database [9] for all experiments. The Weizmann dataset [8] contains 93 video sequences with static background of nine people performing ten natural actions such as running, walking, jacking, waving, etc. with the ground-truth foreground mask. The SFU dataset [9] contains 12 standard video sequences which have dynamic background and complex scenes with the first and second viewing gaze location data by 15 independent viewers. Fixations of the first viewing in the SFU dataset were used as the ground truth.

4.2 Evaluation Metrics

Precision, Recall, and F-Measure. These metrics are used to evaluate performance of the object location detection at a binarized threshold. Similarly to [20], we used an adaptive threshold for each image, which is determined as twice the mean value of salient values over the entire given image.

The F-measure [20] is the overall performance measure computed by the weighted harmonic of precision and recall: $F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 \times Precision + Recall}$. Similarly to [20], we chose $\beta^2 = 0.3$ to weight precision more than recall.

Absolute Correlation Coefficients. The linear Correlation Coefficient (CC) metric [21] focuses on saliency and gaze statistical distributions. To have advantages when comparing average CC from videos, we use Absolute Correlation Coefficient (ACC): $ACC = \left| \frac{\sum_p ((SM(p)-\mu_{SM})(GT(p)-\mu_{GT}))}{\sigma_{SM}\,\sigma_{GT}} \right|$ where SM is the saliency map and GT is the ground truth; $\mu_{SM}, \sigma_{SM}, \mu_{GT}, \sigma_{GT}$ are mean values and standard deviation of SM and GT respectively.

Normalized Scanpath Saliency. The Normalized Scanpath Saliency (NSS) metric [22] focuses on saliency map values at eye gaze positions. This metric quantifies saliency map values at the ground truth locations and normalizes it with saliency variance.

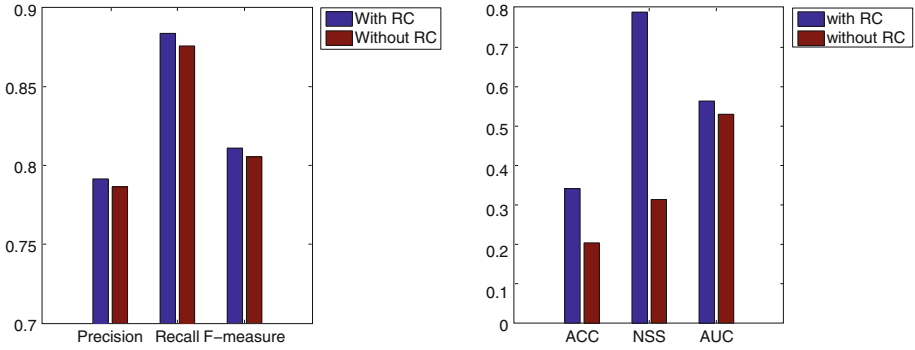
Area Under Curve. In Area Under the Curve (AUC), saliency map is treated as a binary classifier on every pixel where pixels with larger values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated. To reduce influence of the border cut and the center-bias over AUC, we adopted the shuffled-AUC [23], a standard evaluation method used in many recent works. We used an implementation of the shuffled-AUC metric by Z.Bylinskii [24].

5 Evaluation of the Proposed Method

5.1 Evaluation of Introduction to Regional Characteristics

In a video, each pixel value of each frame always changes over time regardless that it belongs to an object or background. Therefore, contrast information using only pixel-based features cannot effectively highlight objects from dynamic background. However, the combination of regional characteristics, derived from region-based features, and contrast information can overcome this limitation because region-based features reduce fluctuation of pixel values in a region. Therefore, our method effectively predict human attention in videos.

To verify this, we conducted experiments to compare Precision, Recall, F-measure values for salient object detection and ACC, NSS, AUC values for eye fixation obtained from our method (denoted by "with RC") with those without regional characteristics (denoted by "without RC"). Results in Fig. 4 indicates that our method outperforms the others in all metrics.



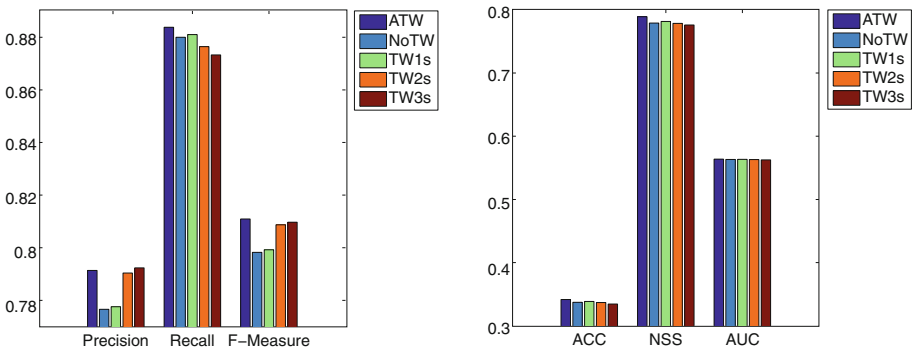
(a) Salient object detection (b) Human eye fixation

Fig. 4. Regional characteristic evaluation. (a) is experimental results for salient object detection on the Weizmann dataset [8]; (b) is experimental results for human eye fixation on the SFU dataset [9].

5.2 Evaluation of Adaptive Temporal Windows

Although combining saliency values of consecutive frames can have positive effect, the employed window size should depend on motion and background. In contrast to the method using a fixed window size, our method adapts the window size to different motion regions in a video. Therefore, our method efficiently utilizes information from consecutive frames to keep temporal consistency between frames.

To verify this, we performed experiments to compare Precision, Recall, F-measure values for salient object detection and ACC, NSS, AUC values for eye fixation obtained from our method (denoted by ATW) with those from the



(a) Salient object detection (b) Human eye fixation

Fig. 5. Adaptive temporal window evaluation. (a) is experimental results for salient object detection on the Weizmann dataset [8]; (b) is experimental results for human eye fixation on the SFU dataset [9].

method without using temporal window (denoted by NoTW), and using a temporal window with the fixed size that corresponds to 1 s, 2 s, and 3 s (denoted by TW1s, TW2s, and TW3s respectively).

Results in Fig. 5 illustrate that our method outperforms the others. Results obtained by the temporal window with a fixed size also perform well, but they can be worse than the ones not using the temporal window when a suitable window size is not chosen.

6 Comparison with State-of-the-Art

6.1 Salient Object Detection

We compared performance of the proposed method with the most recent center-surround contrast based methods using the Weizmann dataset [8]. They are HS [5], LC [25], SAG [26], SO [19], and STS [4]. Among them, LC, SAG, STS are spatial-temporal saliency detection methods, whereas HS and SO are pure spatial methods. We remark that there are some other methods based on the center-surround contrast framework whose results are mostly inferior to the above mentioned methods.

In the first experiment, we used a fixed threshold to binarize saliency maps. In the second experiment, we performed image adaptive binarization of saliency maps. We compared our method with the five methods mentioned above. To evaluate these five methods, we used their publicly available source codes with default configuration set by the authors. Some examples for visual comparison of the methods are shown in Fig. 6, indicating that our method produces the best results on these images.

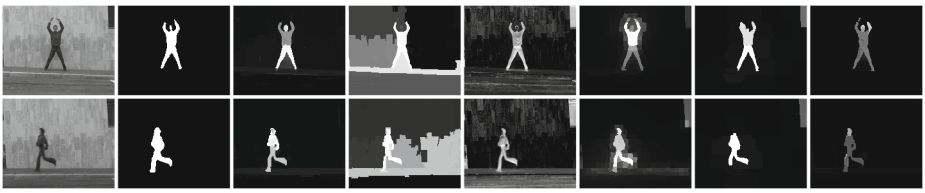


Fig. 6. Visual comparison of our method to the state-of-the-art methods on the Weizmann dataset [8]. From left to right, original images and ground truth are followed by outputs obtained using our method, HS [5], LC [25], SAG [26], SO [19], and STS [4]. Our method achieves the best results.

Image Binarization by a Fixed Threshold. In this experiment, each saliency map is binarized into a binary mask using a saliency threshold θ (θ is changed from 0 to 1). With each θ, the binalized mask is checked against the

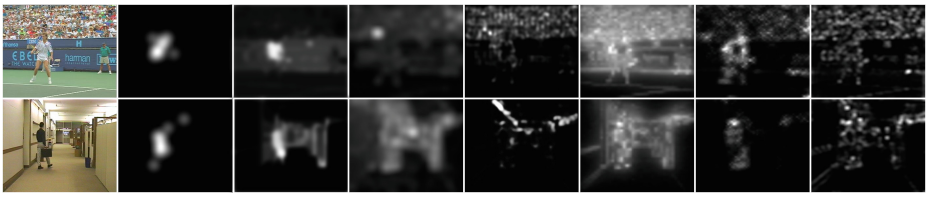
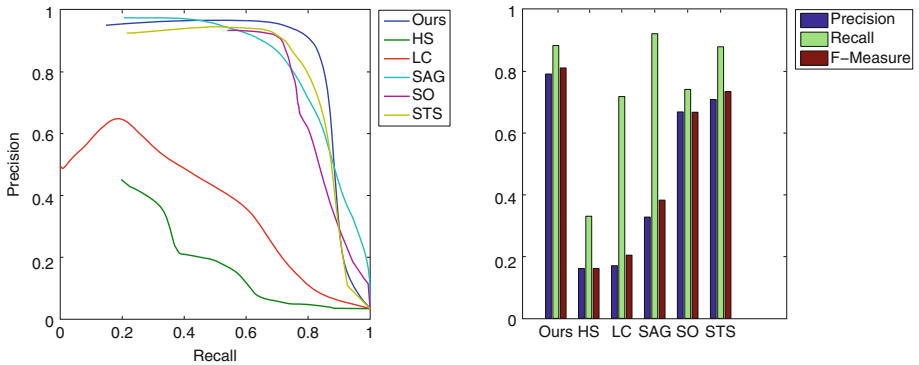


Fig. 7. Visual comparison of our method to the state-of-the-art methods on the SFU dataset [9]. From left to right, original images and ground truth are followed by outputs obtained using our method, BMS [27], DVA [12], RWR [28], SER [10], and SR [29]. Our method achieves the best results.

ground truth to evaluate the accuracy of the salient object detection to compute Precision Recall Curve (PRC) (c.f. Fig. 8 (a)). The PRC is used to evaluate performance of the object location detection because it captures behaviors of both precision and recall under varying thresholds. Therefore, the PRC provides a reliable comparison of how well various saliency maps can highlight salient regions in images.



(a) Fixed threshold (b) Adaptive threshold

Fig. 8. Salient object detection comparison on the Weizmann dataset [8]. (a) is Precision Recall Curves for fixed threshold; (b) is Precision, Recall, and F-Measure for adaptive threshold.

In the PRC, the precision value corresponds to the ratio of salient pixels that are correctly assigned with respect to all the pixels in extracted regions, while the recall value is defined as the percentage of detected salient pixels in relation to the number of salient pixels in the ground truth.gh recall is achieved at the expense of reducing the precision and vice-versa. The results in Fig. 8 (a) show that our method consistently produces saliency maps closer to the ground truth

than the others.This is because the precision value of our method is higher than the others at almost each recall value.

Image Adaptive Binarization. In this experiment, an adaptive threshold depending on obtained saliency for each image is used instead of a fixed threshold. Similarly to [20], the adaptive threshold value is determined as twice the mean value of salient values over a given entire image. Figure 8 (b) shows the Precision, Recall, and F-measure values of our method and the other five methods. Our method outperforms the other methods in all three metrics over the Weizmann dataset.

6.2 Eye Fixation Prediction

We compared performance of our proposed method with dynamic saliency detection methods for human fixation such as BMS [27], DVA [12], RWR [28], SER [10], and SR [29] using the SFU dataset. To evaluate these methods, we used their publicly available source codes with default configuration set by the authors. In order to produce eye fixation maps, our saliency maps are blurred by applying Gaussian blur with zero mean and standard deviation σ (we set $\sigma = 7$). Some examples for visual comparison of the methods are shown in Fig. 7, indicating that our method produces the best results on these images. Figure 9 shows the ACC, NSS, and AUC value comparisons of our method with the other methods. As can be seen, our proposed method outperforms the other methods on all metrics.

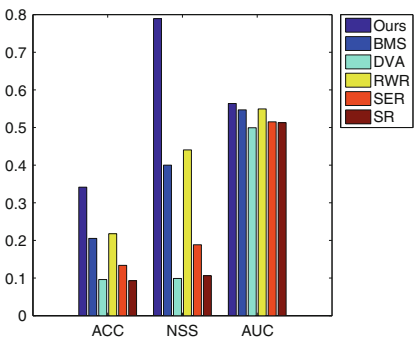


Fig. 9. Eye fixation prediction comparison on the SFU dataset [9], using ACC, NSS, and AUC metrics.

7 Conclusion

In this paper, we present a novel contrast based hierarchical spatial-temporal saliency model for video. Our method effectively integrates pixel-based features and region-based features into a flexible framework so that our method can

utilize both static features and temporal features. Saliency values of consecutive frames are combined by an adaptive temporal window to reduce influence of different motion in a scene, thus the proposed method is robust to dynamic scenes. By introducing region-based features and adaptive temporal window, our method effectively incorporate the dynamic nature of scenes into saliency computation. Experimental results show that our method outperforms state-of-the-art methods on two standard benchmark datasets.

A drawback of the proposed method is not using high-level cues, such as semantic features. Our future work will focus on integration of multiple features and semantic knowledge for further improvement.

References

1. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
2. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE (2009)
3. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1139–1146. IEEE (2013)
4. Zhou, F., Kang, S.B., Cohen, M.: Time-mapping using space-time saliency. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3358–3365, June 2014
5. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155–1162. IEEE (2013)
6. Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., Hiraki, K.: Attention prediction in egocentric video using motion and visual saliency. In: Ho, Y.-S. (ed.) PSIVT 2011, Part I. LNCS, vol. 7087, pp. 277–288. Springer, Heidelberg (2011)
7. Luo, Y., Cheong, L.-F., Cabibihan, J.-J.: Modeling the temporality of saliency. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 205–220. Springer, Heidelberg (2015)
8. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1395–1402. IEEE (2005)
9. Hadizadeh, H., Enriquez, M.J., Bajic, I.V.: Eye-tracking database for a set of standard video sequences. IEEE Trans. on Image Process. **21**(2), 898–903 (2012)
10. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. J. Vision **9**(12), 15 (2009)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
12. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: Advances in Neural Information Processing Systems, pp. 681–688 (2009)
13. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 171–177 (2010)
14. Xiong, C., Xu, C., Corso, J.J.: Streaming hierarchical video segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 626–639. Springer, Heidelberg (2012)

15. Namuduri, K., Mehrotra, R., Ranganathan, N.: Edge detection models based on gabor filters. In: Proceedings of 11th IAPR International Conference on Pattern Recognition, 1992. Vol. III. Conference C: Image, Speech and Signal Analysis, pp. 729–732, August 1992
16. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. MIT, Cambridge (2009)
17. Pele, O., Werman, M.: The quadratic-chi histogram distance family. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 749–762. Springer, Heidelberg (2010)
18. Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. J. Vision **9**(7), 4 (2009)
19. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2814–2821. IEEE (2014)
20. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, pp. 1597–1604. IEEE (2009)
21. Toet, A.: Computational versus psychophysical bottom-up image saliency: a comparative evaluation study. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2131–2146 (2011)
22. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Res. **45**(18), 2397–2416 (2005)
23. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans. Image Process. **22**(1), 55–69 (2013)
24. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: Mit saliency benchmark. http://saliency.mit.edu/
25. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia. pp. 815–824. ACM (2006)
26. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of IEEE CVPR (2015)
27. Zhang, J., Sclaroff, S.: Saliency detection: a Boolean map approach. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 153–160. IEEE (2013)
28. Kim, H., Kim, Y., Sim, J.Y., Kim, C.S.: Spatiotemporal saliency detection for video sequences based on random walk with restart. IEEE Trans. Image Process. **24**(8), 2552–2564 (2015)
29. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2007, pp. 1–8. IEEE (2007)