Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction

Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem
cgds77@gmail.com, {aysunkocak, erkut, aykut}@cs.hacettepe.edu.tr
Department of Computer Engineering, Hacettepe University

Abstract

Predicting where human looks in images has gained a significant popularity in recent years. Compared to the vast number of saliency methods for static images, dynamic saliency estimation remains relatively unexplored. In this work, we propose deep saliency networks based on the two-stream architecture that processes both spatial and temporal information to predict saliency in videos. In particular, we investigate several fusion strategies to combine information coming from spatial and temporal streams and analyze their effectiveness for dynamic saliency prediction. Moreover, to improve the generalization of the saliency networks, we introduce a novel and cognitively grounded data augmentation technique. Experimental results on the DIEM, UCF-Sports datasets show that the proposed approach is able to model human attention mechanism behavior better than the competing methods, achieving state-of-the-art results.

1. Introduction

Humans developed a selective visual attention mechanism that provides an ability to filter out the irrelevant information from a scene and to process a part of a scene instead of the whole. Predicting attention grabbing regions is one of the major problems in computer vision and in that regard computational models for saliency detection have gained increasing popularity lately because of the use in different computer vision problems including image retrieval [9], visual quality assessment [5] video resizing/summarization [2], action recognition [20] and more.

Most of the computational models in the literature are developed to predict salient regions in the static scenes. The previously proposed models aim to detect the regions that are different from their surroundings based on low-level cues like color, orientation, texture, intensity, etc. and/or high-level cues like pedestrians, faces, text, etc. While the low-level features are employed to determine how different a point from its neighborhood, high-level features are employed to guide the saliency detection based on the empirical studies that like humans have a tendency to fixate on certain object classes more than others.

Predicting saliency in videos has more challenges compared to saliency prediction in static images. Since humans have a tendency to focus on moving objects or part of the scenes, the temporal characteristics of the videos have to be considered alongside the spatial characteristics. The first generation of models for video saliency prediction are the extensions of the static saliency models, such as [7, 6, 1, 18]. However, more recent works approach the task from a different point of view and propose much novel solutions [8, 12, 17].

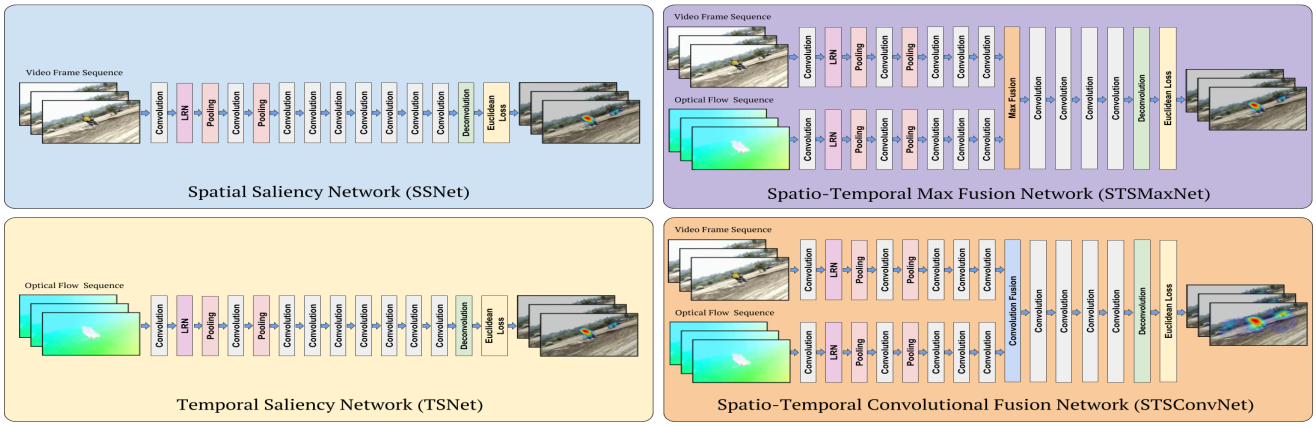
The aim of this study is to investigate saliency for dynamic scenes with deep CNN networks. Our goal is to find an effective model to represent human visual attention mechanism on videos and examine the contributions of the appearance and motion information. In parallel with this aim, we propose a spatio-temporal network to detect saliency in videos. We also propose a new data augmentation technique to boost saliency detection performance for deep models.

2. The Approach

Our spatio-temporal saliency networks for dynamic saliency prediction are based on the so-called two-stream CNNs [19]. Our two-stream models, given in Figure 1(b), integrates both appearance and motion cues with three different methods: by direct averaging, convolution and elementwise max operation. But we first describe our single stream baselines, which are shown in Figure 1(a), that use either appearance or motion information.

2.1. Spatial Saliency Network

The first baseline model is based on finetuning the deep architecture proposed in [15] video. This model uses only appearance cues by considering the RGB video frames as input. The aim is to understand the effect of the spatial information for dynamic saliency. This model is referred as SSNet rest of this paper.



(a) Single stream saliency networks (b) Two-stream saliency networks

Figure 1. (a) The baseline single stream saliency networks. (b) The proposed two-stream spatio-temporal saliency networks.

2.2. Temporal Saliency Network

As mentioned before, human tends to focus on moving local parts (objects, part of an object, etc.) of the scenes and our second baseline is trained to analyze this by using only motion cues for dynamic saliency. The proposed model is same as the SSNet except the input which is optical flow images computed as in [19] from two subsequent frames. We refer this model as TSNet rest of this paper.

2.3. Spatio-Temporal Saliency Networks

Our two-stream network models integrate the spatial and temporal information as shown in Figure 1(b). The biological motivation behind these architectures is the two-streams hypothesis [4] which speculate that human visual cortex is comprised of two distinct streams, namely ventral (what pathway) and the dorsal (where pathway) streams, which are respectively specialized to process appearance and motion information. Some studies argue that dorsal and ventral streams are not strictly independent, but do interact with each other during visual processing. For this reason, we prefer to develop a two-stream model and analyze the effects of interactions between the streams.

In our both of two-stream models, while one stream takes RGB video frames as input, the second one uses corresponding optical flow images. We used different fusion strategies to find the most effective architecture for dynamic saliency prediction:

Fusion via Direct Averaging. This fusion strategy combines spatial and motion information by direct averaging the response maps of the layers before merging. We refer this model as STSAvgNet rest of this paper.

Max Fusion. This model merges two single-stream network via elementwise max fusion. That is, given two feature maps from the spatial and temporal streams, max fusion takes the maximum of these two feature maps at every spatial location. We refer this model as STSMaxNet rest of this paper.

Convolutional Fusion. Our last model fuses spatial and motion streams via a convolution layer. That is, the corresponding feature maps from the spatial and temporal streams are concatenated and then combined with a bank of 1×1 filters. We refer to this network architecture as STSConvNet.

3. Implementation Details

As mentioned before, our single-stream architecture is based on the model that proposed in [15]. The model takes $320 \times 240 \times 3$ pixels images as input and processes them by the operations mentioned in the reference method.

The proposed spatio-temporal architecture consists of two input streams followed by a fusion layer and convolution, deconvolution layers. As shown in Figure 1(b) the single-stream networks resemble each other till fusion layer, which are then combined. We used inputs of size $320 \times 240 \times 3$ pixels for all of our experiments. The optical flow information is extracted via [21] and optical flow images for temporal stream are generated by stacking horizontal and vertical flow components and the magnitude of the flow together.

In our study, we used a different data augmentation strategy instead of the traditional ones like cropping, horizontal flipping or RGB jittering since they alter the visual stimuli used in the eye tracking experiments in collecting the fixation data and we argue that they are not cognitively grounded. In [10], the effects of the resolution on the exploratory behavior of humans are analyzed with an eye-tracking experiment and one of the outcomes of that study is humans are consistent about where they look on low-resolution and high-resolution versions of the same images. Based on this observation, we add our training set the low-resolution versions of the video frames downsampled by

factor 2 and 4. The optical flow images are rescaled to match the down-sampling rate and also we used the fixations obtained from original images as ground truth information.

4. Experimental Results

We employ three different datasets: DIEM (Dynamic Images and Eye Movements) [14], UCF-Sports [13], in our experiments. DIEM dataset consists of eye fixation data of approximately 50 different human subjects for 84 high-definition natural videos including movie trailers, advertisement etc. UCF-Sports dataset consists of 150 videos obtained from 13 different action classes. It is originally used for action recognition but since the eye fixation data is collected for this dataset by Mathe and Sminchisescu [13].

We show the performance of our deep models by comparing state-of-the-art dynamic saliency models, which can be seen in Table 2 and 3. We used several metrics to compare mentioned models with ours: Area Under Curve (AUC), shuffled AUC (sAUC) [22], Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) [16], Normalized Cross Correlation (NCC) and χ^2 distance.

Table 1. sAUC scores of our spatio-temporal saliency networks at different fusion layers on the DIEM dataset.

Fusion Layers	STSMaxNet	STSConvNet
Conv2	0.52	0.71
Conv3	0.67	0.70
Conv4	0.76	0.83
Conv5	0.81	0.84
Conv6	0.80	0.79
Conv7	0.81	0.79

For DIEM dataset, we split the dataset into a training set containing 64 video sequences and a testing set including the remaining 20 representative videos covering different concepts. Specifically, we use the same set of splits used in [17]. As mentioned before, our two-stream networks both have a fusion layer that applies different operations to combine spatial and temporal streams. One of our experiments aims to determine the optimum layers for the fusion process. According to the given sAUC metric results in Table 1, fusing the spatial and temporal streams after the fifth convolution layer achieves the best results for both STSMaxNet and STSConvNet networks. In Figure 2, presents some sample saliency maps extracted with the proposed spatio-temporal saliency networks and the single stream baseline. As can be seen, STSConvNet is the best performing model compared to others. Besides the qualitative results, quantitative results in Table 2 prove that STSConvNet, especially STSConvNet* which is the version of STSConvNet that employs proposed data augmentation performs better compared to other mentioned models and previous dynamic saliency prediction methods. We can

Table 2. Performance comparisons on the DIEM dataset.

	AUC	sAUC	CC	NSS	χ^2	NCC
SSNet	0.72	0.69	0.35	1.85	0.48	0.41
TSNet	0.79	0.77	0.41	1.98	0.40	0.43
STSAvgNet	0.68	0.62	0.37	1.67	0.49	0.37
STSMaxNet	0.83	0.81	0.46	2.01	0.31	0.45
STSConvNet	0.87	0.84	0.47	2.15	0.29	0.46
STSConvNet*	**0.88**	**0.86**	**0.48**	2.18	**0.28**	**0.47**
GBVS [7]	0.74	0.70	0.47	2.04	0.47	0.38
SR [8]	0.69	0.64	0.30	2.22	0.57	0.40
PQFT [6]	0.71	0.67	0.33	1.77	0.52	0.33
Seo-Milanfar [18]	0.59	0.51	0.10	0.12	0.75	0.28
Rudoy *et al.* [17]	–	0.74	–	–	0.31	–
Fang *et al.* [3]	0.71	0.48	0.21	0.55	0.87	0.40
Zhou *et al.* [23]	0.60	0.52	0.13	0.24	0.71	0.22
DWS [11]	0.81	0.79	0.32	**2.97**	0.40	0.39

conclude that applying 1×1 convolutional filters learns the optimal weights to combine appearance and motion feature maps. As shown in Figure 2, our model generates more accurate saliency maps compared to other successful dynamic saliency models, namely GBVS [7], PQFT [6], SR [8], Seo and Milanfar [18], Rudoy et al. [17], Fang et al. [3], Zhou et al. [23], and DWS [11].

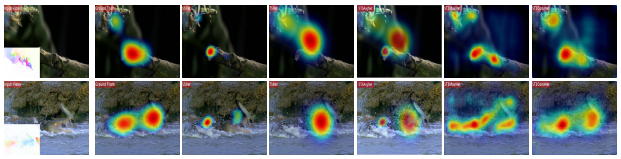


Figure 2. Qualitative evaluation of the proposed saliency network architectures. STSConvNet provides the most accurate prediction as compared to the other network models.

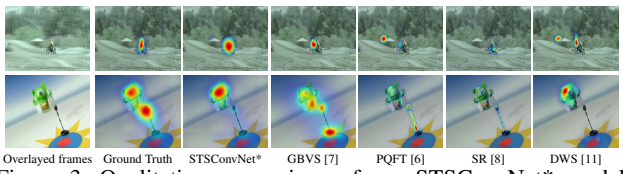


Figure 3. Qualitative comparison of our STSConvNet* model against some dynamic saliency models on DIEM dataset.

To validate generalization ability of our best-performing STSConvNet* model, we perform additional experiments on UCF-Sports dataset. In particular, we do not carry out any training for our model from scratch or finetune it on UCF-Sports but rather use the predictions of the model trained only on DIEM dataset. In Table 3, we provide the performance of our model compared to the previous dynamic saliency models which are publicly available on the web. As can be seen, our STSConvNet* model performs better than the state-of-the-art models according to majority of the evaluation measures. In Figure 4, we can compare our model's results with other dynamic saliency methods.

Table 3. Performance comparisons on the UCF-SPORTS dataset.

	AUC	sAUC	CC	NSS	χ^2	NCC
GBVS [7]	0.83	0.52	0.46	1.82	0.54	**0.59**
SR [8]	0.78	0.69	0.26	1.20	0.42	0.52
PQFT [6]	0.69	0.51	0.29	1.15	0.64	0.48
Seo-Milanfar [18]	0.80	0.72	0.31	1.37	0.56	0.36
Fang *et al.* [3]	**0.85**	0.70	0.44	1.95	0.52	0.33
Zhou *et al.* [23]	0.81	0.72	0.36	1.71	0.56	0.37
DWS [11]	0.76	0.70	0.28	2.01	0.40	0.49
STSConvNet*	0.82	**0.75**	**0.48**	**2.13**	**0.39**	0.54

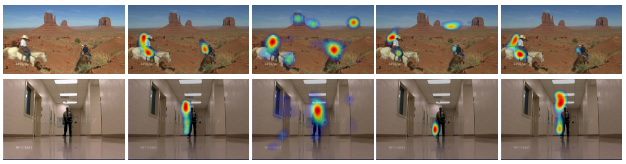


Figure 4. Qualitative comparison of our STSConvNet* model against some previous dynamic saliency models on UCF-Sports dataset. Our spatio-temporal saliency network outperforms the others.

5. Conclusion

In this work, we have investigated several deep architectures for predicting saliency in dynamic scenes. Our proposed spatio-temporal saliency networks are built based on two-stream architecture and employ different fusion strategies, namely direct averaging, max fusion and convolutional fusion, to integrate appearance and motion features, and they are all trainable in an end-to-end manner. To train these saliency networks more effectively, we also propose an task-specific and cognitively grounded data augmentation strategy that utilizes low-resolution versions of the video frames and the ground truth saliency maps. Our experimental results demonstrate that the proposed STSConvNet model achieves superior performance over the state-of-the-art methods on DIEM and UCF-Sports datasets.

References

[1] X. Cui, Q. Liu, and D. Metaxas. Temporal spectral residual: fast motion saliency detection. In *ACM MM*, pages 617–620, 2009.

[2] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[3] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Trans. Image Processsing*, 23(9):3910–3921, 2014.

[4] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

[5] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, 18(6):1098–1110, 2016.

[6] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, pages 1–8, 2008.

[7] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[8] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8, 2007.

[9] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang. Integrating visual saliency and consistency for re-ranking image search results. *IEEE Transactions on Multimedia*, 13(4):653–661, 2011.

[10] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *Journal of Vision*, 11(4):14–14, 2011.

[11] V. Leboran, A. Garcia-Diaz, X. Fdez-Vidal, and X. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

[12] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*, pages 842–856, 2012.

[13] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 2015.

[14] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[15] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016.

[16] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, 2005.

[17] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, pages 1147–1154, 2013.

[18] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 2009.

[19] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[20] X. Wang, L. Gao, J. Song, and H. Shen. Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition. *IEEE Signal Processing Letters*, 24(4):510–514, 2017.

[21] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013.

[22] L. Zhang, M. H. Tong, T. K. M. anf H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.

[23] F. Zhou, S. B. Kang, and M. F. Cohen. Time-mapping using space-time saliency. In *CVPR*, 2014.