SIGNBOARD SALIENCY DETECTION IN STREET VIDEOS

Onkar Krishna, Kiyoharu Aizawa
Dept. of Information and Communication Engineering
The University of Tokyo, Japan

Saskia Reimerth[*]
Technische Universitt Wien, Austria

ABSTRACT

During the last few decades researchers in computer vision have proposed various saliency models for images with the common goal of classifying the image content by using the measure of importance. However, compared to still images, there is only a limited number of saliency detection algorithms proposed for video signals. However, predicting where a person looks in a video is relevant for applications such as advertisement design, video re-targeting and editing. In this work, we propose a novel method for video saliency detection that aims to detect the relative ranking of saliencies of signboards in street videos. For that reason, we collected eye-gaze data of participants viewing various street videos in free viewing and task viewing scenarios, where the task was to identify a place to have lunch at. Further, we quantitatively analyzed the collected eye-gaze data in order to generate the relative ranking of the signboards in the free viewing and the task viewing scenario. Based on the analysis' results, we propose a video saliency detection algorithm which can more accurately predict the relative saliencies of signboards in street videos. It can be seen that the prediction accuracy of our proposed model outperforms the existing video saliency detection algorithms.

Index Terms— saliency, computational modeling, free viewing, task viewing, heat map

1. INTRODUCTION

The human visual system has developed the ability to process a scene by selecting the most relevant parts of the scene unconsciously. This mechanism is called selective attention, and has been developed in order to quickly spot danger, which was key to human survival. Only within the last 10-15 years, researchers in computer vision exploited the concept of human selective attention for computational models of saliency prediction. However, the saliency models developed so far are limited to predict conspicuous location in static images [1-5], whereas only a few saliency detection algorithms are proposed for video signals [6-8]. Accurately predicting salient regions in video signals is critical in many video processing applications such as object detection, video re-targeting, robot navigation, and video compression. However, it is still undiscovered to build a practical model for saliency detection which can accurately mimic the human visual system.

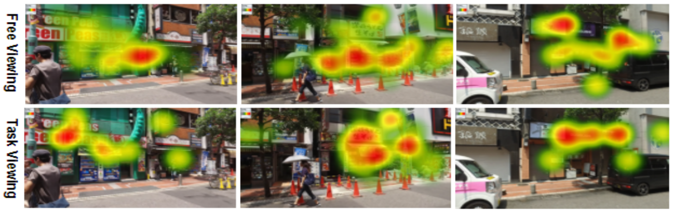


Fig. 1: Visualization of different tendencies of gaze landings around the signboards in free viewing and task viewing scenario.

Visual attention in a scene is either being ”pulled” to a particular location by bottom-up features of the scene such as color, intensity, and orientation, or ”pushed” to a particular location by the top-down factors such as given tasks and goals. There are several studies in vision research that have revealed that gaze distribution during scene viewing is highly dependent on the given task (e.g. free viewing and task viewing) [9-11]. However, most of these studies have been conducted on static images, and the impact of cognitive factors (e.g. a given task, intentions) on video signals is largely ignored [14-15]. This study quantitatively analyzes how different viewing strategies (free viewing and task viewing) influence the gaze landings around the signboards in a street video, and then propose an algorithm which can predict more accurately the relative saliency of signboards during free viewing and task viewing.

Most of the saliency detection algorithms are based on the Feature Integration Theory (FIT) [12] of which the main idea is to compute bottom-up features (spatial and temporal) in parallel and to fuse their saliency to get a so-called master saliency map. However, some of the saliency detection algorithms are based on the guided search model [13] where the master saliency map is generated by combining the saliency map obtained from top-down information together with bottom-up saliency maps. Following the guided search model, most of the recent video saliency detection algorithms are based upon integrating the spatial information with the temporal information [6-8]. The main difference between algorithms lies in the way bottom-up and top-down information is represented. The study reported in [8] integrated top-down

[*]The third author performed the work while at visiting student at The University of Tokyo

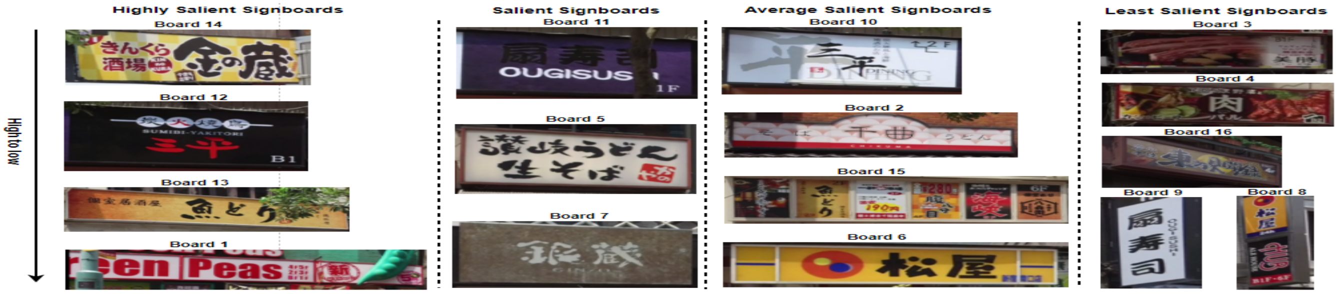


Fig. 2: *Segmented signboards of restaurants in the street video and their corresponding saliency rankings, generated from the eye-gaze data collected over thirty participants, (task viewing ranking: board 13, 5, 14, 11, 15, 12, 2, 1, 6, 7, 10, 16, 3, 8, 4, 9.)*

information represented as semantic cues, such as faces and speech, with bottom-up saliency maps obtained from classical saliency algorithms. Similarly, the study in [17] represented bottom-up information as a color histogram of the video frames, while the temporal map calculated the motion between images by applying RANSAC.

Although there are many advances in video saliency detection algorithms being made recently, how successfully these algorithms can be used in different computer vision applications still remains a key issue. The performance of these applications depends on the prediction accuracy of the saliency detection algorithm. The proposed study analyzes how well a bottom-up saliency prediction algorithm can be used in a novel application of predicting the relative ranking of the saliency of the signboards in street videos.

Our main contribution is three-folded. First, we introduce a new eye gaze dataset collected over 30 observers viewing two street videos in both free viewing and task viewing scenarios, the tasks being to look for a place to have either lunch at. Secondly, we propose a metric for quantitative analysis of the collected eye gaze data to find differences in tendencies of gaze distribution for signboards in street videos during the free viewing and task viewing. Finally, we propose a modification to an existing algorithm to improve the prediction accuracy of signboard saliency in street videos.

2. EXPERIMENT AND ANALYSIS

2.1. Participants and Video Stimuli

A total of 30 participants attended the experiment including 15 university students (3 female, 12 male, age range 21-31, mean age 24.1) and 15 (4 female, 11 male, age range 66-80, mean age 73.1) elderly subjects recruited from retirement job centers in Tokyo. All the subjects reported normal or corrected to normal vision. Participants were non-familiar with the experimental procedure and had not seen the video stimuli before the experiment. All of the participants signed a consent form before the start of the experiment.

Two different street videos, each with a duration of two minutes thirty seconds, were used in the experiment. Each video consisted of 4,473 frames in total. Tobii x2-60 eye-tracker was used for recording the eye-gaze data, whereas the fixations and saccades were detected by the default Tobii fixation filter (I-VT fixation filter). Video stimuli were displayed at a 20 inch LCD monitor of 40 cm width, and all the participants were viewing the stimuli at a 65cm distance from the monitor surface.

2.2. Procedure and Task

The 30 participants who took part in the study were divided into two groups of 15 participants each. In order to avoid repeating the same video in free viewing and task viewing mode for any one participant, one group of the participants watched the video in free viewing mode, whereas the another group watched the same video with the given task of finding a place to have lunch at. Each trial started with the gaze calibration followed by the eye tracking while viewing the videos. Before the video stimuli began, participants were instructed to either view freely or to fulfill the task.

2.3. Analysis

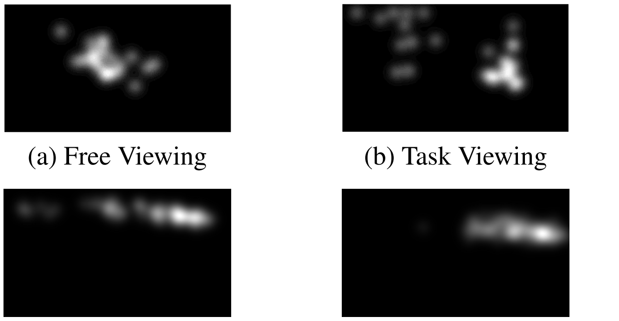
The different tendencies of the gaze landings during free viewing versus task viewing can be seen in the heat maps shown in Figure 1. Further, in order to determine the relative saliencies of the signboards of restaurants in the street video during the free viewing and task viewing mode, the distribution of fixation locations around the signboards has been analyzed. To perform this study, first, we manually labeled two instances of each signboard appearing in the video and then interpolated the label for the rest of the frames containing the same signboard. There was a total of 16 signboards labeled in the street video for the full duration of their appearance (Figure 2). Further, two different approaches were adopted to analyze the gaze distribution around the signboards labeled in the previous step. First, we simply measured the total number of gazes that have landed on each signboard during the free viewing and the task viewing scenario for the whole duration of the signboard's appearance. For the second one, we developed explorativeness metrics to measure the differences in the scene exploration tendency during the free viewing and the task viewing scenario.

The relative ranking of the signboard saliencies based on the gaze counting during free viewing is shown in Figure 2. One way ANOVA was conducted to measure the statistical difference between the signboard's saliencies in the highly salient, salient, average salient and least salient category. A significant difference was reported for gaze landings among the signboards belonging to the four different categories, $F(3, 29) = 3.56, p < 0.001$.

Further, entropy-based metrics were developed to measure the explorativeness during free viewing and task viewing. In order to do that, we first generated saliency maps by convolving a Gaussian similar to [3] over all fixation locations recorded for every second (30 frames) into a single frame. As an output of this step, we have generated 149 saliency maps (total 4473 frames divided by 30), showing the area explored during each second of the video in a single frame. As shown in Figure 3(a, b), the average of all these 149 maps shows the average rate of exploration during free viewing and task viewing of the street videos. Finally, the explorativeness is measured by measuring the entropy of the 149 saliency maps. Formulation of the explorativeness is as follows:

$$H(I_j) = \sum_{l} h_{I_j}(l) * log(L \, / \, h_{I_j}(l)) \qquad (1)$$

where I_j is the saliency map of the total gazes recorded during one second of viewing for which entropy is calculated and $h_{I_j}(l)$ is the histogram entry of intensity value l in image I_j, and L is the total number of pixels in I_j.



(a) Free Viewing (b) Task Viewing

(c) Board 14-Free Viewing (d) Board 14-Task Viewing

Fig. 3: (a, b) Average tendency of the rate of video exploration during free viewing and task viewing. (c, d) Different tendency of explorativeness around board 14 only during free or task viewing

The average score of the entropy value suggests higher explorativeness during task viewing than free viewing (task viewing - 1.94, free viewing - 1.40). The one-way ANOVA showed an effect of a task in scene exploration tendencies, $F(1, 148) = 22.13, p < 0.001$. Similarly, we have measured the explorativeness tendency for the labeled locations of restaurant signboards only (Figure 3(c, d)), where the results showed comparable explorativeness in both viewing scenarios. We can see from Figure 3 (a, b), that the center bias tendency is very different during free viewing and task viewing.

Table 1: Prediction accuracy of different algorithms for the gaze over the full duration for free viewing and task viewing

	GBVS	s_map	m_map	e_map	Itti
Free	0.7969	0.7626	0.6996	0.7429	0.7961
Task	0.7717	0.7350	0.6836	0.7045	0.7483

The center bias for two different viewing modes can further be measured by measuring the euclidean distance between the centroid of the average maps (Figure 3(a, b)) and the center pixel of the image. The higher euclidean distance for task viewing suggests the lower center bias in the task viewing scenario compared to the free viewing scenario (203 and 267 are the euclidean distances in pixels for free viewing and task viewing).

The proposed analysis suggests three major findings that can be applied in upgrading the existing video saliency algorithm to improve the prediction performance for the relative rankings of signboard saliencies. First, we generated the ground truth rankings of the signboards based on the eye-gaze data collected in the free viewing and the task viewing scenario. Secondly, the explorativeness results indicate a higher exploration tendency during task viewing compared to free viewing. Lastly, we discovered a higher center-bias for free viewing than task viewing.

3. PROPOSED MODEL FOR SIGNBOARD SALIENCY DETECTION IN FREE VIEWING

Before proposing any algorithm for signboard saliency detection, we evaluated the performance of existing video saliency algorithms in predicting the gazes landed on the signboards. We selected a few state-of-the art algorithms [1-2] with their motion included version, which can be aligned with the goal of signboard saliency detection in street videos. The performance of these algorithms are evaluated to answer the following two questions: (1) How accurately can they predict the gaze points landed anywhere within the frame during free viewing and task viewing? (2) How well can these algorithms predict the gazes only for the signboards in those two viewing scenarios?

As shown in Table 1 the performance of [1] and [2] in predicting the gazes for the whole frame is better than the s_map [18], m_map[19], and e_map [6]. Similarly, the result of prediction accuracy for signboards only showed that GBVS [2] and Itti's [1] perform better in predicting signboard saliencies for the least and highest salient signboards as labeled in Table 2. Motivated by the prediction accuracy of Itti's model, especially for the signboards belonging to the least and highest salient categories, we applied the recommendations from the analysis' results to upgrade Itti's model with motion features [1] to predict the signboard saliencies more accurately for free viewing and task viewing scenarios. Another reason for selecting Itti's model was based on the fact that most exist-

Table 2: The signboard saliency scores (AUC score) for the lowest salient signboards (determined by gaze data) generated by different algorithms in free viewing and task viewing.

Board Ranking (Task)		12	15	13	11	Avg.
Board Ranking (Free)		14	16	13	12	
Board Name		3	8	9	16	
GBVS[2]	Free	0.73	0.53	0.65	0.70	0.65
	Task	0.76	0.63	0.52	0.70	0.65
Itti's[1]	Free	0.79	0.61	0.56	0.63	0.64
	Task	0.71	0.72	0.57	0.71	0.67
s_map[18]	Free	0.67	0.61	0.68	0.60	0.64
	Task	0.69	0.65	0.58	0.77	0.67
m_map[19]	Free	0.61	0.65	0.61	0.61	0.62
	Task	0.63	0.62	0.54	0.56	0.58
e_map[6]	Free	0.70	0.62	0.54	0.61	0.61
	Task	0.65	0.56	0.48	0.60	0.57
Ours	Free	0.70	0.51	0.49	0.63	0.58
	Task	0.71	0.55	0.45	0.56	0.56

Table 3: The signboard saliency scores (AUC score) for the highest salient signboards (determined by gaze data) generated by different algorithms in free viewing and task viewing (higher score is better).

Board ranking (Task)		8	7	1	4	Avg.
Board ranking (Free)		4	5	3	1	
Board Name		1	2	13	14	
GBVS[2]	Free	0.79	0.87	0.75	0.76	0.79
	Task	0.83	0.82	0.64	0.64	0.73
Itti's[1]	Free	0.82	0.79	0.88	0.84	0.83
	Task	0.82	0.75	0.71	0.71	0.74
s_map[18]	Free	0.81	0.74	0.88	0.79	0.80
	Task	0.75	0.72	0.66	0.68	0.70
m_map[19]	Free	0.66	0.67	0.57	0.64	0.63
	Task	0.69	0.62	0.61	0.57	0.62
e_map[6]	Free	0.76	0.72	0.75	0.74	0.74
	Task	0.77	0.72	0.61	0.61	0.67
Ours	Free	0.81	0.86	0.79	0.81	0.81
	Task	0.83	0.84	0.73	0.75	0.78

ing bottom-up models follow the same basic architecture proposed by Itti et al. In these models the following basic structure can be observed: (a) Basic visual features such as color, intensity, orientation and motion are extracted over multiple scales of the image, where each scale represents a different level of detail in the scene. (b) All features are investigated in parallel, to obtain the conspicuity map for each feature channel. (c) These features are integrated to obtain the saliency map.

The analysis' result suggested that participants explored the video more during task viewing than the free viewing. Thus, to make our model adapt to free viewing and task viewing scenarios, we focused on a feature scale selection mechanism, where we identified the subsets of the feature map scales that best represented the different levels of details viewed by the observers during free viewing and task viewing. Further, we illustrated step combining feature maps at different scales to generate the final saliency map.

$$Intensity = \bigoplus_{i=s}^{6} \mathcal{N}(Intensity_i) \tag{2}$$

$$Color = \bigoplus_{i=s}^{6} [\mathcal{N}(\mathcal{RG}_i) + \mathcal{N}(\mathcal{BY}_i)] \tag{3}$$

$$Orientation = \sum_{\theta \in \{0,45,90,135\}} \bigoplus_{i=s}^{6} \mathcal{N}(Orientation_i(\theta)) \tag{4}$$

where $\mathcal{N}$ represents the normalization and s is the starting index from where maps were taken, scale 1 (finer) to scale 6 (coarser). Similarly to the motion features, we have selected a subset of scales for representing the explorativeness in the two different viewing scenarios.

The experimental result shows that the prediction accuracy of the signboard saliency detection during free viewing is highest for the following subset of the coarser scales $s = 4, 5, 6$, conversely the task viewing prediction accuracy improved for the finer scales $s = 1, 2, 3$. This result is in accordance with our previous findings of the explorativeness' differences in free viewing and task viewing. Further, we tuned the center bias weights in the existing model, based on the analysis' results that free viewing has a higher center bias than task viewing. As shown in Table 2 and Table 3, the weight tuning for the center bias together with the subset selection for explorativeness in free viewing and task viewing have slightly improved the relative saliency prediction for both highly and least salient signboards .

4. CONCLUSION

This paper presents a novel application of video saliency detection for ranking signboards within a street video based on the relative signboard saliencies. The major contributions of this work are a collection of eye-gaze data for 2 street videos for both free viewing and task viewing scenarios, and further, the proposal of a quantitative analysis method based on the rate of the explorativeness and center bias metrics. Finally those results were used in upgrading the basic saliency model for predicting signboard saliencies more accurately for free viewing and task viewing.

5. ACKNOWLEDGMENT

We are thankful to VTEC laboratories Inc. for providing the video used in this study. We are also thankful to Dr. Go Irie, NTT Corporation for providing valuable insights.

6. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 20, 11 (1998), 12541259.

[2] Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency. In Advances in neural information processing systems. 545552.

[3] Tilke Judd, Krista Ehinger, Frdo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In Computer Vision, 2009 IEEE 12th international conference on. IEEE, 21062113.

[4] Erkut Erdem and Aykut Erdem. 2013. Visual saliency estimation by nonlinearly integrating features using region covariances. Journal of vision 13, 4 (2013), 1111.

[5] Josselin Gautier and Olivier Le Meur. 2012. A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. Cognitive Computation 4, 2 (2012), 141156.

[6] Fang, Y., Wang, Z., Lin, W., and Fang, Z. (2014). Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Transactions on Image Processing, 23(9), 3910-3921.

[7] Guo, C., and Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE transactions on image processing, 19(1), 185-198.

[8] Zhai, Y., and Shah, M. (2006, October). Visual attention detection in video sequences using spatiotemporal cues. In Proceedings of the 14th ACM international conference on Multimedia (pp. 815-824). ACM.

[9] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, Eye movements in iconic visual search, Vision Research, Vol.42, no. 11, pp.l447-1463, 2002.

[10] C. Balkenius, Attention, habituation and conditioning: Toward a computational model, Cognitive Science Quarterly, vol.1, pp.171-214, 2000.

[11] J. Taylor and N. Fragopanagos, Modelling the interaction of attention and emotion, Proc. International Joint Conference on Neural Networks (IJCNN), pp. 1663-1668, 2005.

[12] A. M. Treisman, The perception of features and objects. In Attention: Selection, awareness, and control, A. Baddeley and L. Weiskrantz, Eds. Clarendon Press, Oxford, pp.5-35. 1993.

[13] J. M. Wolfe, Guided search 2.0: A revised model of visual search, Psychonomic Bulletin and Review 1, 2, pp.202-238, 1993.

[14] Onkar Krishna, Kiyoharu Aizawa, Andrea Helo, and Rama Pia. 2017. Gaze Distribution Analysis and Saliency Prediction Across Age Groups. arXiv preprint arXiv:1705.07284 (2017).

[15] Krishna, O., and Aizawa, K. (2017, September). Age-adapted saliency model with depth bias. In Proceedings of the ACM Symposium on Applied Perception (p. 5). ACM.

[16] Le Meur, O., and Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behavior research methods, 45(1), 251-266.

[17] Y. Ma, X. Hua, L. Lu, and H. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Trans. Multimedia, vol. 7, no. 5, pp. 907919, Oct. 2005.

[18] Yuming Fang, Zhenzhong Chen, Weisi Lin, Chia-Wen Lin: Saliency Detection in the Compressed Domain for Adaptive Image Retargeting. IEEE Transactions on Image Processing 21(9): 3888-3901 (2012)

[19] Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin, A Video Saliency Detection Model in Compressed Domain. IEEE Trans. Circuits Syst. Video Techn. 24(1): 27-38, 2014.