Can Saliency Map Models Predict Human Egocentric Visual Attention?

Kentaro Yamada[1], Yusuke Sugano[1], Takahiro Okabe[1]
Yoichi Sato[1], Akihiro Sugimoto[2], and Kazuo Hiraki[3]

[1] The University of Tokyo, Tokyo, Japan, 153-8505
{yamada,sugano,takahiro,ysato}@iis.u-tokyo.ac.jp
[2] National Institute of Informatics, Tokyo, Japan, 101-8430
sugimoto@nii.ac.jp
[3] The University of Tokyo, Tokyo, Japan, 153-8902
khiraki@idea.c.u-tokyo.ac.jp

Abstract. The validity of using conventional saliency map models to predict human attention was investigated for video captured with an egocentric camera. Since conventional visual saliency models do not take into account visual motion caused by camera motion, high visual saliency may be erroneously assigned to regions that are not actually visually salient. To evaluate the validity of using saliency map models for egocentric vision, an experiment was carried out to examine the correlation between visual saliency maps and measured gaze points for egocentric vision. The results show that conventional saliency map models can predict visually salient regions better than chance for egocentric vision and that the accuracy decreases significantly with an increase in visual motion induced by egomotion, which is presumably compensated for in the human visual system. This latter finding indicates that a visual saliency model is needed that can better predict human visual attention from egocentric videos.

1 Introduction

Our visual focus of attention is an important clue for inferring our internal state and therefore can be used effectively for developing human-centric media such as interactive advertising, intelligent transportation systems, and attentive user interfaces. Since our visual focus of attention is closely related to our gaze, many gaze sensing techniques based on various approaches have been developed. However, it is still a difficult task to measure our gaze in unconstrained settings.

An alternative way of estimating the visual focus of attention is to use a visual saliency map model. Inspired by psychological studies of visual attention [1], Koch and Ullman proposed the concept of the saliency map model [2]. Itti et al. subsequently proposed a computational model [3] for predicting which image locations attract more human attention. Since then, many types of saliency map models have been proposed [4–9]. The models have been applied not only to static images but also to video clips by incorporating low-level dynamic image

features such as motion and flicker [4]. Studies based on actual gaze measurement [10–12] have demonstrated that such saliency maps match distributions of actual human attention well. However, those studies considered only recorded images and videos. The saliency maps were computed from images shown to human subjects, and their effectiveness was evaluated against the gaze coordinates on the display. While such visual saliency map models can be used for certain applications such as image editing, they lack an important aspect: consideration of the visual motion caused by motion of the observer, i.e., visual motion seen in a static scene captured by a moving camera.

Egocentric vision refers to a research field analyzing dynamic scenes seen from egocentric perspectives, e.g., taken from a head-mounted camera. Egocentric perspective cameras are well suited for monitoring daily ego activities. Accurate prediction of visual attention in egocentric vision would prove useful in various fields, including health care, education, entertainment, and human-resource management. However, the mechanism of visual attention naturally differs significantly in egocentric perspectives. For instance, visual stimuli caused by egomotion are compensated for in egocentric vision, but such a mechanism is not considered in conventional saliency map models. Since conventional models have not been examined for egocentric videos, whether they are valid for such videos is unclear. We have investigated the validity of using conventional saliency map models for egocentric vision. Egocentric videos were captured using a head-mounted camera, and gaze measurements were made using a wearable gaze recorder. The performances of several saliency models and features were quantitatively determined and compared, and the characteristics of human attention in egocentric vision were discussed. To the best of our knowledge, this is the first experimental evaluation of the performance of saliency map models for egocentric vision.

2 Related Work

In this section, we first introduce background theory on visual saliency and briefly review previous work on computational saliency map models.

Due to a person's limited capacity to process incoming information, the amount of information to be processed at a time must be limited. That is why a mechanism of attention is needed to efficiently select and focus on an important subset of the available information [13]. The same holds true for the human visual system; visual attention is necessary to enable a person to handle the large amount of information received through the eyes.

A key to understanding the mechanism of visual attention is feature integration theory [1]. The human visual system first divides incoming images into simple visual features [14]. Since natural objects usually have two or more features, after processing each simple feature separately, the visual system reintegrates the incoming image information. Treisman et al. concluded from their studies that the human mechanism of visual attention includes integration of such visual cues. On the basis of this theory, Koch and Ullman proposed the concept of a visual

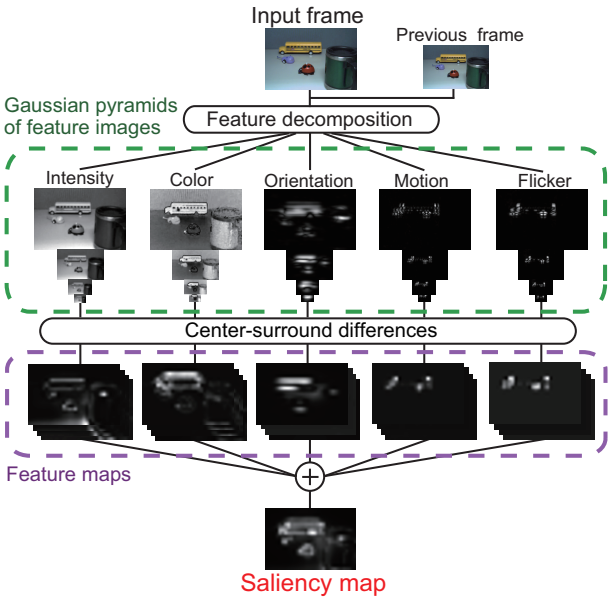


Fig. 1. Procedure for computing saliency maps for videos

saliency map: a two-dimensional topographic map that encodes saliency values for a scene [2]. Those values are generated by integrating simple visual features, and they represent how strongly the region attracts a person's attention.

Itti et al. [3] proposed and developed a fully bottom-up computational saliency map model. They introduced procedures for extracting simple visual features from images; the saliency values are computed through procedures for imitating visual receptive fields. The input is static images, and the output is saliency maps corresponding to the input images. The model was later extended by adding two dynamic features, motion and flicker, so that it can deal with dynamic scenes [4].

Other approaches to saliency map modeling have been proposed. For instance, in the recently introduced graph-based approach [7–9], graph representations of input images are generated by defining dissimilarity functions and distance functions between nodes. Saliency values are computed through steady-state analysis of the graphs. Studies using this approach focused mainly on the procedures for computing the saliency values from simple image features rather than on validating the efficiency of the image features used in the models.

3 Procedure for Computing Saliency Maps for Videos

In this study, we used two representative saliency map models to evaluate the validity of using saliency map models for egocentric vision. One is Itti et al.'s model [4] which is based on the center-surround mechanism, and the other is

Harel et al.'s graph-based model [7]. We first introduce the computational procedure of Itti et al.'s model, and then explain Harel et al.'s model.

Figure 1 illustrates the procedure which consists of three main stages. In the first stage, feature decomposition generates Gaussian pyramids of *feature images* from an input frame. In the second stage, "center-surround" mechanism generates *feature maps* from feature images; i.e., saliency maps are computed from each feature. In the third stage, the feature maps are normalized and integrated into a single saliency map.

In the first stage, the input image is decomposed into five types of visual feature images using simple linear filters. The features are typically intensity, color and orientation as static features, and motion and flicker as dynamic features. The intensity feature image is obtained as the average of the red, green, and blue channels of the input images. Itti et al. used two difference images generated by sets of two color channels, i.e., red-green and blue-yellow, for the color feature images. In contrast, we use the Derrington-Krauskopf-Lennie (DKL) color space [15] as color features instead of these difference images. The DKL color space is defined physiologically by three channels used for color processing in the retina and thalamus. Orientation feature images are computed from the intensity image using four oriented $(0°, 45°, 90°, 135°)$ Gabor filters.

Two input frames are required for obtaining flicker and motion feature images. The flicker feature image is computed from the absolute difference between the intensity feature images in the current and previous frames. The motion feature images are obtained from the spatially shifted differences between every four orientation feature images of the current and previous frames. As a result, 12 feature images are obtained: one for intensity, two for color, four for orientation, one for flicker, and four for motion. Next, nine spatial scales (scale zero = 1:1 to scale eight = 1:256) are created using dyadic Gaussian pyramids [16] for each feature image.

In the next stage, feature maps are computed from these Gaussian pyramids using the center-surround mechanism. We made six sets of two different sizes of Gaussian pyramids. Six feature maps were computed from each feature image using across-scale image subtraction, which is obtained by interpolation to the finer scale and point-wise subtraction.

In the last stage, the final saliency map is obtained by combining the 72 normalized feature maps (six for intensity, 12 for color, and 24 for orientation, six for flicker, 24 for motion). The normalization is performed by globally multiplying each feature map by $(M - \bar{m})$, where M is the map's global maximum and $\bar{m}$ is the average of its other local maxima. This normalization process suppresses the feature maps with more peaks and thus enhances the feature maps with fewer peaks.

Harel et al.'s model [7] follows the graph-based approach in the second and the third stages. The feature maps and final saliency map are generated by computing the equilibrium distributions of Markov chain graphs. For the second stage, they defined a dissimilarity function and a distance function between nodes and multiplied them together to obtain the weight of each node. In the

(a) (b)

Fig. 2. (a) EMR-9 [17], mobile eye tracking system developed by NAC Image Technology. EMR-9 has two eye cameras and two IR light sources to measure gaze movements at 240 [Hz]. It captures egocentric video at 30 [fps] using head-mounted scene camera. Horizontal view angle of scene camera is 121°, and resolution of recorded video is 640×480. (b) Example video frame captured the scene camera during experiment.

last stage, they obtained the weight of each node by multiplying the value of the location on the feature maps by the distance function.

4 Experiment

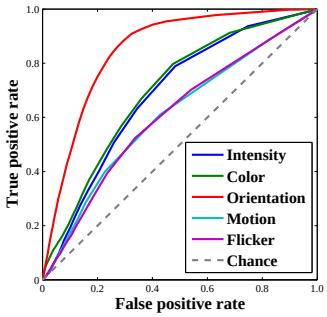
As summarized above, conventional saliency map models use simple, low-level image features as sources to compute saliency maps. They are designed to compute visual saliency for recorded images and videos, but no consideration is given to dealing with visual motion induced by camera motion. To evaluate the validity of using conventional models for egocentric vision, we conducted an experiment.

4.1 Experimental Procedure

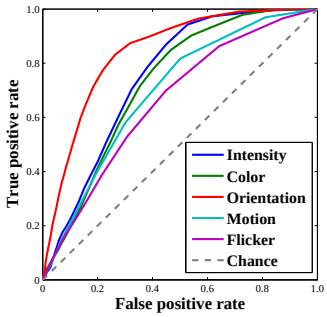
To enable us to evaluate the validity of saliency map models for egocentric vision, we designed an experiment that would enable us to determine the correlation between the actual gaze points and the saliency maps for videos captured with a head-mounted camera.

We used the EMR-9 mobile eye tracking system developed by NAC Image Technology [17] to determine the gaze points and to capture egocentric videos. As shown in Figure 2(a), the EMR-9 has two eye cameras and two IR light sources for measuring gaze movement at 240 [Hz]. The scene camera attached to the head captures egocentric video at 30 [fps]. The horizontal view angle of the scene camera was 121°, and the resolution of the recorded video was 640×480.

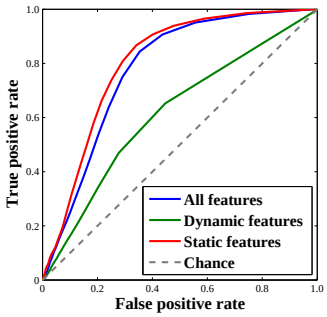
We used the saliency map models of Itti et al. [4] and Harel et al. [7] as baseline models. The experiment was conducted in a room. Four human subjects (one at a time) sat on a chair while another person walked randomly around the room. The subjects were asked to look around the room by moving their head freely for one minute. Figure 2(b) shows an example video frame captured by



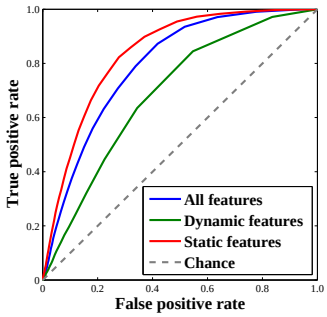
(a) Each feature/Itti et al. [4]



(b) Each feature/Harel et al. [7]



(c) Combined features/Itti et al. [4]



(d) Combined features/Harel et al. [7]

Fig. 3. ROC curves for each feature ((a) Itti et al. [4], (b) Harel et al. [7]) and for static, dynamic, and all features ((c) Itti et al. [4], (d) Harel et al. [7]). Curves were calculated by changing saliency threshold values from minimum to maximum. Horizontal axis indicates false positive rate, i.e., rate of pixels above threshold. Vertical axis indicates true positive rate, i.e., rate of gaze points for which saliency value of corresponding point on saliency map was higher than threshold.

the scene camera. We obtained about 12,000 gaze points for each subject after removing errors caused by eye blinks.

Human attention is affected by performing a task, but the high-level mechanism of attention cannot be treated efficiently with conventional saliency map models. Since the purpose of our study was to examine the validity of saliency map models for egocentric vision, and thus, we did not assign a task to the subjects.

4.2 Results

To examine how each feature contributes to the accuracy of estimating attention, we compared the correlation between each *feature saliency map*, computed using only one feature, and the actual gaze points. The curves in Figure 3 are the average receiver operating characteristic (ROC) curves, which were calculated

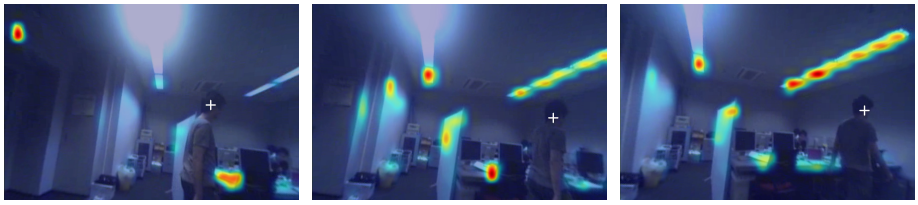


Fig. 4. Examples of gaze trajectory of subject facing moving object (walking person). Images are overlaid with motion feature saliency maps. Crosses show gaze points.

by changing the saliency threshold values from minimum to maximum. The horizontal axis indicates the false positive rate, i.e., the rate of pixels on the map above a threshold. The vertical axis indicates the true positive rate, i.e., the rate of gaze points for which the saliency value of the corresponding point on the saliency map was higher than the threshold.

Figures 3 (a) and (b) compare the feature saliency maps explained in Section **3**. Figure 3 (a) shows the results of using Itti et al.'s model [4], and Figure 3 (b) shows the results of using Harel et al.'s model [7]. Figures 3 (c) and (d) compare the static, dynamic, and standard saliency maps. The static maps were computed using only the static features (intensity, color, and orientation), and the dynamic maps were computed using only dynamic features (motion and flicker). The standard maps were computed using all the features. Figure 3 (c) shows the results of using Itti et al.'s model [4], and Figure 3 (d) shows the results of using Harel et al.'s model [7]. The areas under the curves (AUC) of these three curves, a measure of prediction performance, are shown in Table 1. These results indicate that these saliency map models can predict human egocentric visual attention better than chance. However, with both models, the dynamic features did not contribute to performance. In fact, they even reduced accuracy.

Table 1. AUC of combined saliency maps for two models.

	Static features	Dynamic features	All features
Itti et al. [4]	0.803	0.615	0.778
Harel et al. [7]	0.838	0.690	0.793

4.3 Discussion

Our experimental results show that the performance of the dynamic features, motion and flicker, significantly degrades prediction performance for egocentric

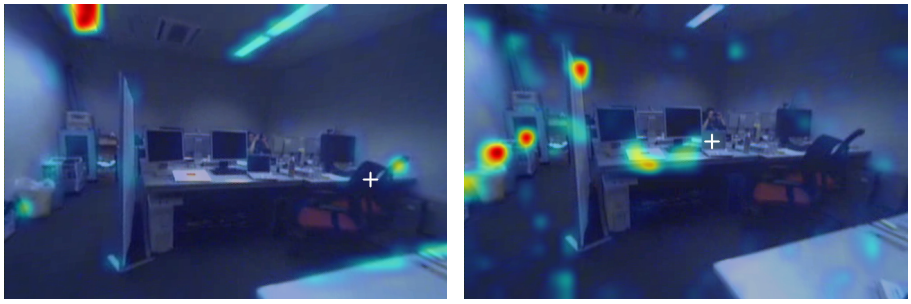
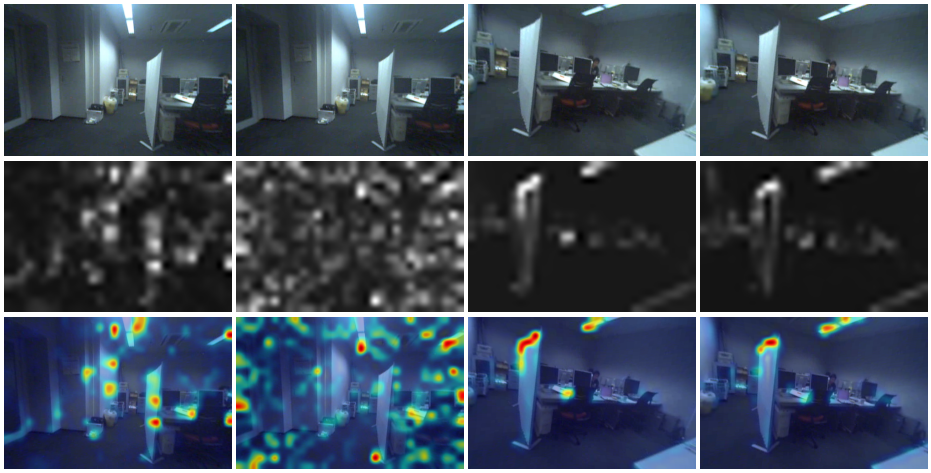


Fig. 5. Example of the scene in which object quickly changed its color (laptop monitor). Images are overlaid with flicker feature saliency maps. Crosses show gaze points.

vision. However, during the experiment, we observed situations in which dynamic visual stimuli attracted the subject's attention. Figure 4 shows examples of the gaze trajectory when the subject was facing a moving object (walking person). Figure 5 shows an example scene in which an object quickly changed its color (laptop monitor). In these cases, the subject paid attention to the dynamic changes of the visual features; however, large saliency values are given to the other locations which did not dynamically change. Hence, previously proposed features could not capture dynamic visual stimuli appropriately in our experimental situation.

Unlike the case with recorded images and videos, the case in which we are interested includes the effects of egomotion. While human beings have the ability to compensate for egomotion [18], conventional saliency map models do not have a mechanism for such compensation, so high saliency values appear in dynamic feature saliency maps regardless of whether they are caused by egomotion.

Figure 6 shows example dynamic feature saliency maps with and without the effect of egomotion. Figure 6 (a) and (b) shows video frames with small egomotion, and (c) and (d) show ones with large egomotion. Figure 6 (a) and (c) are motion feature saliency maps, and (b) and (d) are flicker feature saliency maps. The images in the top row are input images, those in the middle row are feature saliency maps, and those in the bottom row are input images overlaid with feature saliency maps. As shown in Figures 6 (a) and (b), many peaks appear within dynamic feature saliency maps when the egomotion is small. Since they are suppressed by the normalization in the last combining step, explained in Section **3**, these peaks do not substantially affect the final saliency map. In contrast, as shown in Figures 6 (c) and (d), large saliency values are given to the locations with large disparity and to the edges of large intensity difference caused by large egomotion. These feature saliency maps can greatly affect the final saliency map. This indicates that, to model dynamic visual stimuli efficiently, it is necessary to compensate for large egomotion.



(a)Small egomotion, motion feature map | (b)Small egomotion, flicker feature map | (c)Large egomotion, motion feature map | (d)Large egomotion, flicker feature map

Fig. 6. Examples of dynamic feature saliency maps with and without effect of egomotion. (a) and (b) show video frames with small egomotion, and (c) and (d) show frames with large egomotion. (a) and (c) are motion feature saliency maps, and (b) and (d) are flicker feature saliency maps. Images in the top row are input images, those in middle row are feature saliency maps, and those in bottom row are input images overlaid with feature saliency maps.

5 Conclusion and Future Work

We have investigated the validity of using saliency maps computed from videos captured from an egocentric perspective by experimentally examining the correlation between saliency maps and gaze points. The results show that saliency map models can predict human egocentric visual attention better than chance; however, the dynamic features decreased their performance for egocentric vision because these models cannot model the way a person compensates for the effects of egomotion. The models thus need to be improved to enable them to deal with egocentric videos.

We plan to conduct more experiments under various conditions, e.g., in outdoor scenes and with walking subjects. We also plan to develop a motion compensation mechanism so that the dynamic feature maps work better for egocentric vision.

References

1. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12** (1980) 97–136
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology **4** (1985) 219–227

3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259
4. Itti, L., Dhavale, N., Pighin, F., et al.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: SPIE 48th AnnualInternational Symposiumon Optical Science and Technology. Volume 5200. (2003) 64–78
5. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 693–708
6. Cerf, M., Harel, J. and Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. Advances in Neural Information Processing Systems **20** (2008) 241–248
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. Advances in Neural Information Processing Systems **19** (2006) 545–552
8. Costa, L.: Visual saliency and atention as random walks on complex networks. ArXiv Physics e-prints (2006)
9. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring visual saliency by site entropy rate. In: Computer Vision and Pattern Recognition, IEEE (2010) 2368–2375
10. Foulsham, T., Underwood, G.: What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision **8** (2008) 1–17
11. Itti, L.: Quantitative modelling of perceptual salience at human eye position. Visual Cognition **14** (2006) 959–984
12. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research **42** (2002) 107–123
13. Ward, L.M.: Attention. Scholarpedia **3** (2008) 1538
14. Broadbent, D.: Perception and communication. Pergamon Press (1958)
15. Derrington, A., Krauskopf, J., Lennie, P.: Chromatic mechanisms in lateral geniculate nucleus of macaque. The Journal of Physiology **357** (1984) 241–265
16. Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., Anderson, C.: Overcomplete steerable pyramid filters and rotation invariance. In: Computer Vision and Pattern Recognition, IEEE (1994) 222–228
17. nac Image Technology Inc.: EMR-9. http://www.nacinc.com/products/Eye-Tracking-Products/EMR-9/ (2008)
18. Howard, I.: The optokinetic system. The Vestibulo-ocular Reflex and Vertigo (1993) 163–184