Attention Prediction in Egocentric Video Using Motion and Visual Saliency

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. We propose a method of predicting human egocentric visual attention using bottom-up visual saliency and egomotion information. Computational models of visual saliency are often employed to predict human attention; however, its mechanism and effectiveness have not been fully explored in egocentric vision. The purpose of our framework is to compute attention maps from an egocentric video that can be used to infer a person's visual attention. In addition to a standard visual saliency model, two kinds of attention maps are computed based on a camera's rotation velocity and direction of movement. These rotation-based and translation-based attention maps are aggregated with a bottom-up saliency map to enhance the accuracy with which the person's gaze positions can be predicted. The efficiency of the proposed framework was examined in real environments by using a head-mounted gaze tracker, and we found that the egomotion-based attention maps contributed to accurately predicting human visual attention.

Keywords: Visual saliency, visual attention, first-person vision, camera motion estimation.

1 Introduction

Visual attention can be an important cue to infer the internal states of humans. Techniques to predict human visual attention have been employed in various applications in the area of, e.g., attentive user interfaces and interactive advertisements. One of the most direct ways of inferring visual attention is to measure the human gaze [7]; however, it is still a difficult task to measure our gaze in casual and unconstrained settings.

An alternative way of estimating the visual focus of attention is to use a visual saliency map model. Inspired by psychological studies on visual attention [24], Koch and Ullman proposed the concept of the saliency map model [17]. Itti et al. subsequently proposed a computational model [15] of visual saliency to identify image regions that attract more human attention. Following their study, many types of saliency map models have been proposed through the years [14,1,2,8,3,25]. Studies using gaze measurements [5,12,20] have also demonstrated that the saliency maps agree well with actual distributions of human attention.

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 277–288, 2011.

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 suited for monitoring daily ego-activities, and hence accurate predictions of egocentric visual attention will be useful in various fields including health care, education, entertainment, and human-resource management. There has been much work on video attention analysis [18,21,13]; however, methods of analyzing egocentric visual attention have yet to be sufficiently explored. Saliency maps in these studies were computed from images shown to human subjects using monitors, and their effectiveness was evaluated against the gaze points given on the monitors. Hence, it still remains an unresolved question as to how we can predict visual attention accurately in egocentric videos that include visual motions caused by human head motion.

We propose a new framework in this paper to compute attention maps from egocentric videos using bottom-up visual saliency and egomotion information. Two kinds of egomotion-based attention maps, i.e., rotation-based and translation-based maps are computed in our framework and they are aggregated with the bottom-up saliency maps to produce accurate attention maps.

Camera motion has been employed to analyze attention in home videos [18,21]. Intentional human head motion in egocentric videos can have a stronger relationship with attention directed. Hillair et al. proposed a method of predicting egocentric visual attention in virtual reality environments based on the rotation factor of head movement [10,11]. Fukuchi et al. discussed the effect that focus of expansion (FOE) of moving pictures had in attracting human attention and they provided some experimental evaluations of FOE-enhanced saliency maps [6]. Although the basic idea behind our work was similar to that in these studies, we applied the framework to real egocentric scenes and motion-based maps were computed purely using input video without requiring additional sensors.

It is a well-known fact that humans tend to look at the center of images and a simple centering bias map can also contribute to enhancing the accuracy of saliency maps [16]. Our proposed attention maps can be seen as improved centering bias maps that are well-suited to egocentric vision. The effect of using motion-based attention maps is examined in a real setting using a mobile gaze tracker, and a comparison with a centering map is also discussed in Section 3.

2 Prediction of Visual Attention Using Saliency and Egomotion

The goal of this work was to predict visual attention by only using an egocentric video. Fig. 1 outlines the flow for our proposed framework. While bottom-up visual saliency maps are computed from input egocentric video, motion maps are computed using a person's egomotion. These additional motion maps are integrated into the visual saliency maps, and the resulting map achieves higher accuracy in predicting human attention. Details on the computations for the visual saliency maps and the motion maps are described in the following sections.

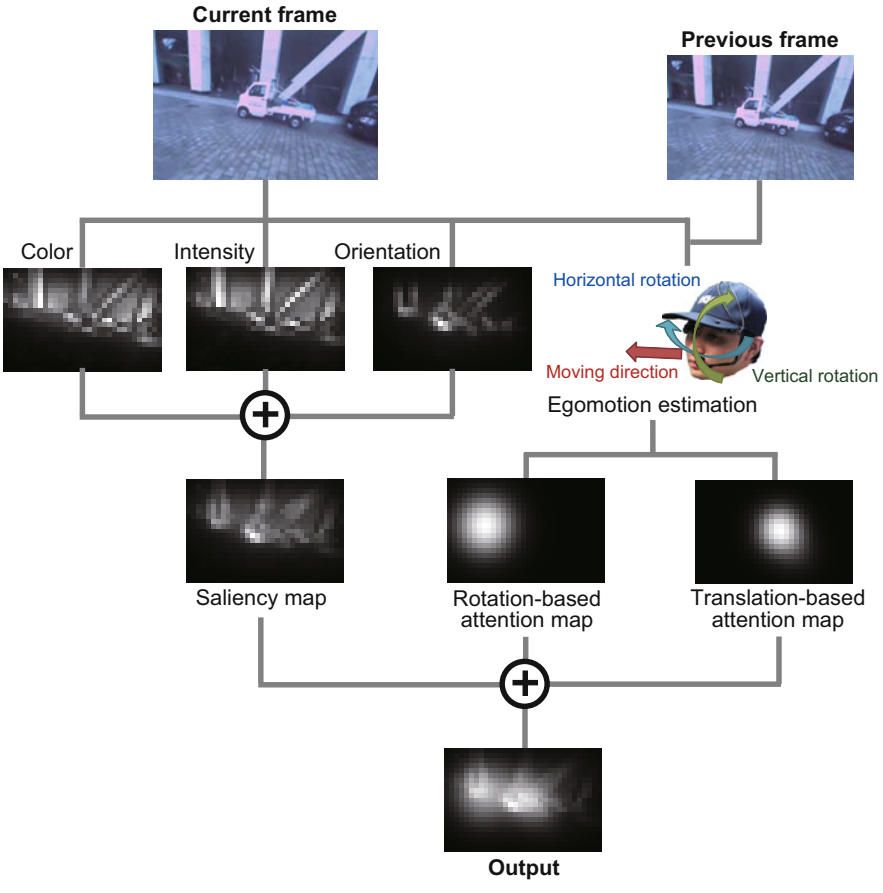


Fig. 1. Flow for our proposed framework. While bottom-up visual saliency maps are computed from input egocentric video, motion maps are computed using person's egomotion. These additional motion maps are integrated into visual saliency maps, and resulting map achieves higher accuracy in predicting human attention.

2.1 Computation of Visual Saliency Maps

We used the graph-based visual saliency (GBVS) model proposed by Harel et al. [8] in this work to compute the bottom-up saliency maps. Since it has previously been reported that saliency maps using dynamic features (motion and flicker) reduce the accuracy of saliency maps in egocentric scenes [26], we only employed static features, i.e., color, intensity and orientation to compute the saliency maps. As discussed above, the core concept in computational visual saliency is extracting regions with vastly different image features than their surrounding regions. Saliency maps in the GBVS model are generated by computing the equilibrium distributions of Markov chain graphs. Graphs are defined with nodes corresponding to pixels, and higher transition probabilities are assigned between dissimilar nodes (=pixels). Higher values are given in this way to

nodes with distinctive image features in their equilibrium distribution and these can be used as saliency maps. Readers should refer to [8] for more details. Saliency maps are computed from each of the three features, and combined with equal weights to generate the final saliency map.

2.2 Computation of Attention Maps from Egomotion

Motion-based attention maps were computed using a person's egomotion in addition to the above visual saliency maps. We employed two kinds of attention maps in this work: rotation-based and translation-based. The computation consisted of three steps: 1) we estimated camera motion from the egocentric video, 2) estimated angular velocity and generating rotation-based attention maps, and 3) estimated the direction of movement and generated translation-based attention maps. We assumed that the camera's intrinsic parameters were known in this work and the lens distortion would be corrected through calibration. The camera was also assumed to be attached to the person's head so that its coordinates were identical to his/her visual field. Details on the three steps are described in what follows.

Estimation of Camera Motion. First, camera motion between two consecutive frames was computed using epipolar geometry, and rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$ were obtained. Feature flows between the two frames were acquired using the Kanade-Lucas-Tomasi feature tracker [23,22], and an eight-point algorithm [9] was then applied to compute the fundamental matrix, $\boldsymbol{F}$. RANSAC [4] was used to robustly select the eight points without being affected by outliers caused by items such as moving objects. Since the intrinsic parameters were known, $\boldsymbol{R}$ and $\boldsymbol{t}$ could be obtained from $\boldsymbol{F}$.

Rotation-Based Attention Map. The rotation angle around each axis was computed from $\boldsymbol{R}$ in the second step, and the rotation-based attention map was generated using horizontal and vertical angular velocities. Let us denote the horizontal and vertical axes of the egocentric video as x and y, the camera's optical axis as z, and the rotation angles around these axes as $\theta_x, \theta_y, \theta_z$. Since it is assumed that the camera and the person's visual field share the same coordinates, the horizontal and vertical rotation angles of the head correspond to θ_y and θ_x. Given rotation matrix $\boldsymbol{R}$ and if we assume a x-y-z rotation order, θ_x and θ_y can be uniquely determined (θ_z is set to 0 if $\theta_y = \pm\frac{\pi}{2}$). By denoting the frame rate of the video as f [fps], the horizontal and vertical angular velocities can be written as $\omega_x = 180 f \theta_x / \pi$ and $\omega_y = 180 f \theta_y / \pi$.

We drew a 2-D Gaussian circle based on the angular velocities with a fixed variance to generate rotation-based attention maps. Hillair et al. [10] reported a strong correlation between gaze positions and angular velocities when the velocity was less than about $100[\mathrm{deg}/s]$. With larger velocity, Gaze positions tend to be almost fixed. According to their report, we define the center of the Gaussian (x, y) as illustrated in Fig. 2:

$$x = \begin{cases} \frac{\omega_y}{100} \cdot \frac{w}{k} & (|\omega_y| \leq 100) \\ \frac{w}{k} & (\omega_y > 100) \\ -\frac{w}{k} & (\omega_y < -100) \end{cases} \tag{1}$$

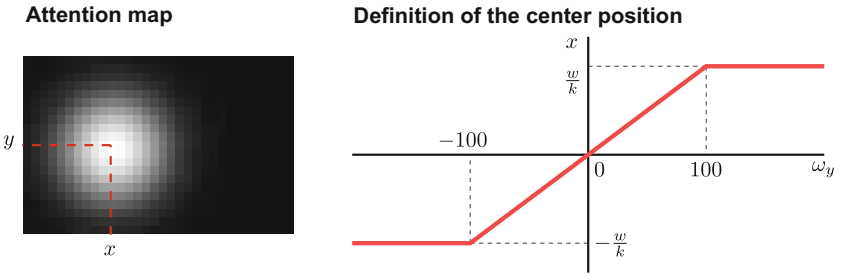


Fig. 2. Rotation-based attention map. 2-D Gaussian circle is drawn with fixed variance to generate rotation-based attention maps based on angular velocities. According to report by Hillair et al. [10], center of Gaussian is defined so that it is proportional to angular velocity within the range of $100[\mathrm{deg}/s]$.

and

$$y = \begin{cases} -\frac{\omega_x}{100} \cdot \frac{h}{l} & (|\omega_x| \leq 100) \\ -\frac{h}{l} & (\omega_x > 100) \\ \frac{h}{l} & (\omega_x < -100), \end{cases} \tag{2}$$

where w, h indicate the width and height of the attention map and k, l are parameters according to the camera's angle of view.

Translation-Based Attention Map. Another attention map is generated in the third step based on the direction of the person's movement. The FOE of the input visual stimuli during translatory movements indicates the direction of movement. Similarly to [6], we generate the motion-based attention map based on the assumption that surrounding regions of the FOE attract more attention. We calculate the FOE of the input video as follows.

Egocentric videos usually contain independently moving objects and the person can also perform rotational movements. Therefore, the intersecting point of their feature flows does not always correspond to the FOE as illustrated in Fig. 3. We first rejected feature flows in this work that were identified as outliers when computing fundamental matrix $\boldsymbol{F}$ and only inlier flows were used in further processing.

Next, the rotational and translational components of the flow were separated. Let us denote the current image as $\mathrm{I}^{(t)}$ and the previous image as $\mathrm{I}^{(t-1)}$. If we can rotate $\mathrm{I}^{(t-1)}$ using the previously computed rotation matrix, $\boldsymbol{R}$, the relationship between the rotated image, $\mathrm{I}_R^{(t-1)}$, and I can be described by the translation vector, $\boldsymbol{t}$. If we denote the camera's intrinsic matrix as $\boldsymbol{A}$, pixel coordinates $\boldsymbol{m}^{(t-1)}$ and $\boldsymbol{m}_R^{(t-1)}$ of the feature point in $\mathrm{I}^{(t-1)}$ and $\mathrm{I}_R^{(t-1)}$ can be written in homogeneous coordinates as

$$\boldsymbol{m}^{(t-1)} \sim \boldsymbol{A}\boldsymbol{x}^{(t-1)}, \tag{3}$$

$$\boldsymbol{m}_R^{(t-1)} \sim \boldsymbol{A}\boldsymbol{x}_R^{(t-1)},, \tag{4}$$

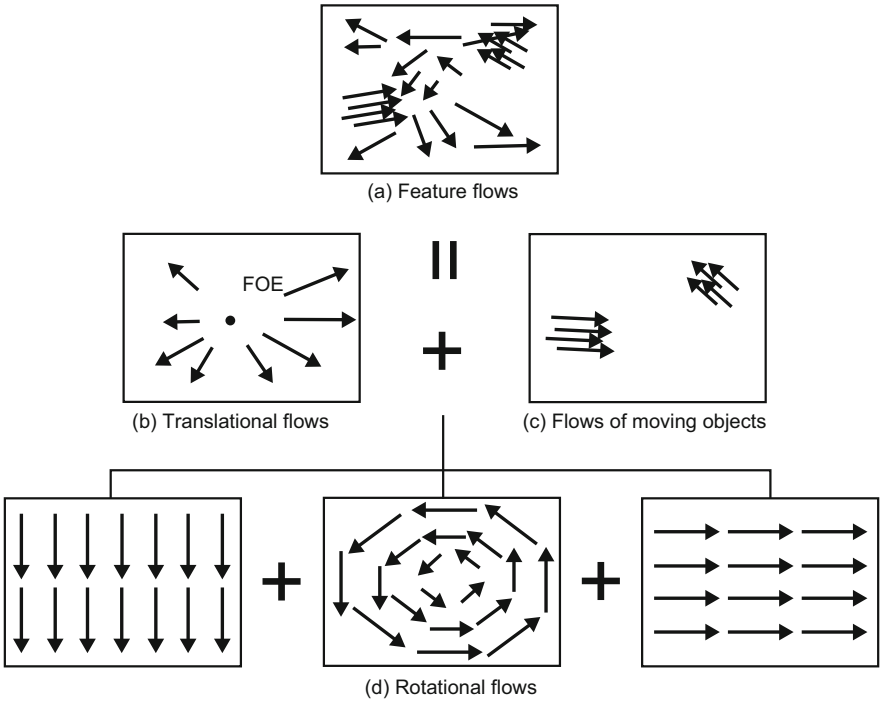


Fig. 3. Components of feature flows. Intersection point of translational flow (b) corresponds to Focus of Expansion (FOE) and it indicates direction of camera movement. However, feature flows computed from egocentric video (a) include flows caused by independently moving objects (c) and rotational movements (d). Components corresponding to (c) and (d) must first be separated from computed flow (a) to estimate FOE of input frame.

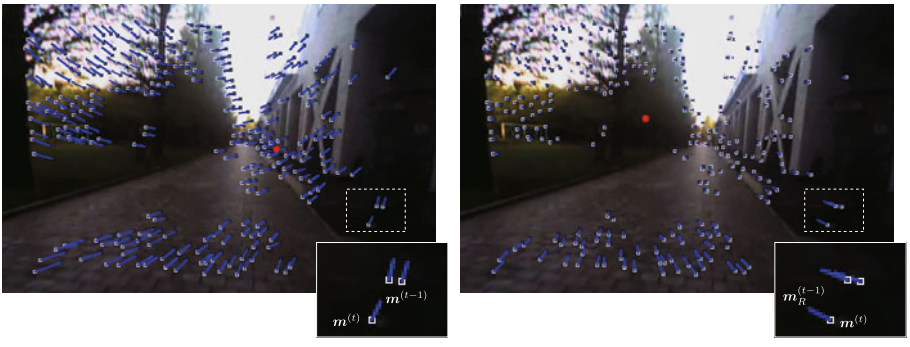
where $\boldsymbol{x}^{(t-1)}$ and $\boldsymbol{x}_R^{(t-1)}$ indicate the normalized image coordinates of the feature point. As discussed above, the following relationship also holds:

$$\boldsymbol{x}_R^{(t-1)} \sim \boldsymbol{R}\boldsymbol{x}^{(t-1)}, \tag{5}$$

and hence $\boldsymbol{m}_R^{(t-1)}$ can be written as

$$\boldsymbol{m}_R^{(t-1)} \sim \boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^{-1}\boldsymbol{m}^{(t-1)}. \tag{6}$$

By applying Eq. (6) to all coordinates $\boldsymbol{m}^{(t-1)}$ of inlier flows, the translational components of flow $\boldsymbol{m}^{(t)} - \boldsymbol{m}_R^{(t-1)}$ can be computed. The FOE is computed as the point with the minimum Euclid distance to all the translational flows. Fig. 4 shows a example of all feature flows and the separated translational flows. The bright rectangles indicate feature positions in current frame $\boldsymbol{m}^{(t)}$, and the dark rectangles indicate feature positions in original image $\boldsymbol{m}^{(t-1)}$ (a) and rotated image $\boldsymbol{m}_R^{(t-1)}$. The circles overlaid in the images indicate the computed FOE.



(a) Raw feature flows (b) Translational flows

Fig. 4. Separation of translational flows. By rotating previous frame using rotation matrix $\boldsymbol{R}$, translational flows (b) can be obtained from raw feature flows (a). Each image shows current frame $\mathrm{I}^{(t)}$. Bright rectangles indicate feature positions in current frame $\boldsymbol{m}^{(t)}$, and dark rectangles indicate feature positions in original image $\boldsymbol{m}^{(t-1)}$ (a) and rotated image $\boldsymbol{m}_R^{(t-1)}$. Circles overlaid in images indicate FOEs that are computed as point with minimum Euclid distance to all translational flows.

The above process computes the FOE based only on two successive frames; however, using multiple video frames will lead to more accurate computation of the moving direction. For this reason, we computed the FOEs between all K pairs of $\mathrm{I}^{(t)}$ and $\mathrm{I}^{(t-k)}$ ($k = 1, 2, \ldots, K$, and $K = 15$ in this work). A motion-based attention map is generated from the K FOEs by Gaussian kernel density estimation.

2.3 Aggregation of Maps

The bottom-up visual saliency maps and the egomotion-based attention maps are then aggregated to compute the final attention map. All maps are summed with equal weights, and the summed map is then normalized to have fixed maximum and minimum values. Fig. 5 shows some examples of visual saliency maps, attention maps, and the final aggregated map. We evaluated three combinations of the maps in this work: A) saliency + rotation + translation, B) saliency + rotation, and C) saliency + translation. This is further discussed in Section 3.

3 Experiments

Here, we describe the details on the experiments we carried out to evaluate what effect using motion-based attention maps had. We used a head-mounted gaze tracker to capture real egocentric videos and ground-truth gaze points. The prediction accuracy of the attention maps was assessed with the receiver operating characteristic (ROC) curves of the maps similarly to evaluating visual saliency maps. The prediction accuracy of our maps was also compared with a simple centering bias map to further demonstrate the efficiency of our method.

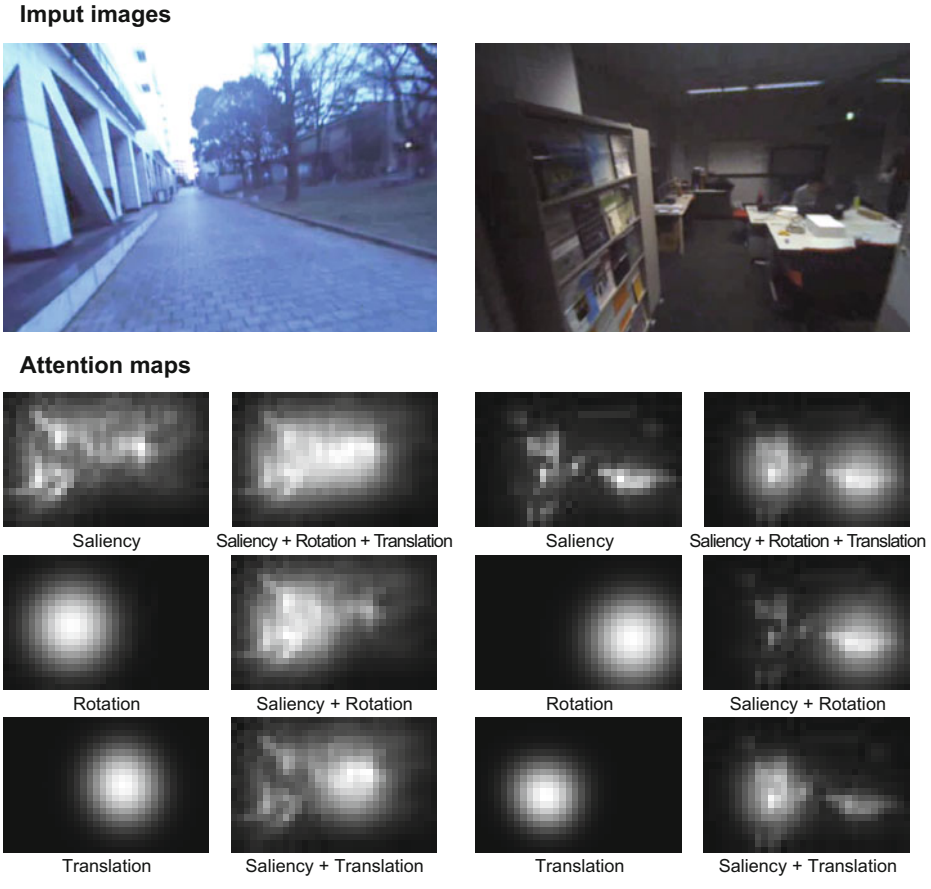
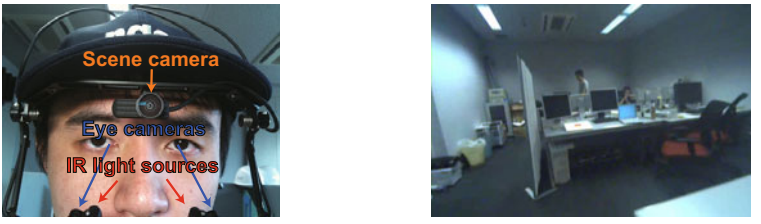


Fig. 5. Examples of attention maps. Top row shows input images, and other images show examples of visual saliency maps (saliency), motion-based attention maps (rotation and translation), and three different types of their combinations.

3.1 Experimental Settings

A mobile gaze tracker, the EMR-9 [19] developed by NAC Image Technology, was used in the experiments. A scene camera was installed on the EMR-9 as seen in Fig. 6(a), and it captured egocentric video of the subject at 30 [Hz]. The horizontal field of view of the scene camera was $121°$, and the resolution of the egocentric video was 640×480 [pixels]. EMR-9 also had two eye cameras and two infrared light sources, and recorded the ground-truth gaze points on the egocentric video at 240 [Hz].



(a) (b)

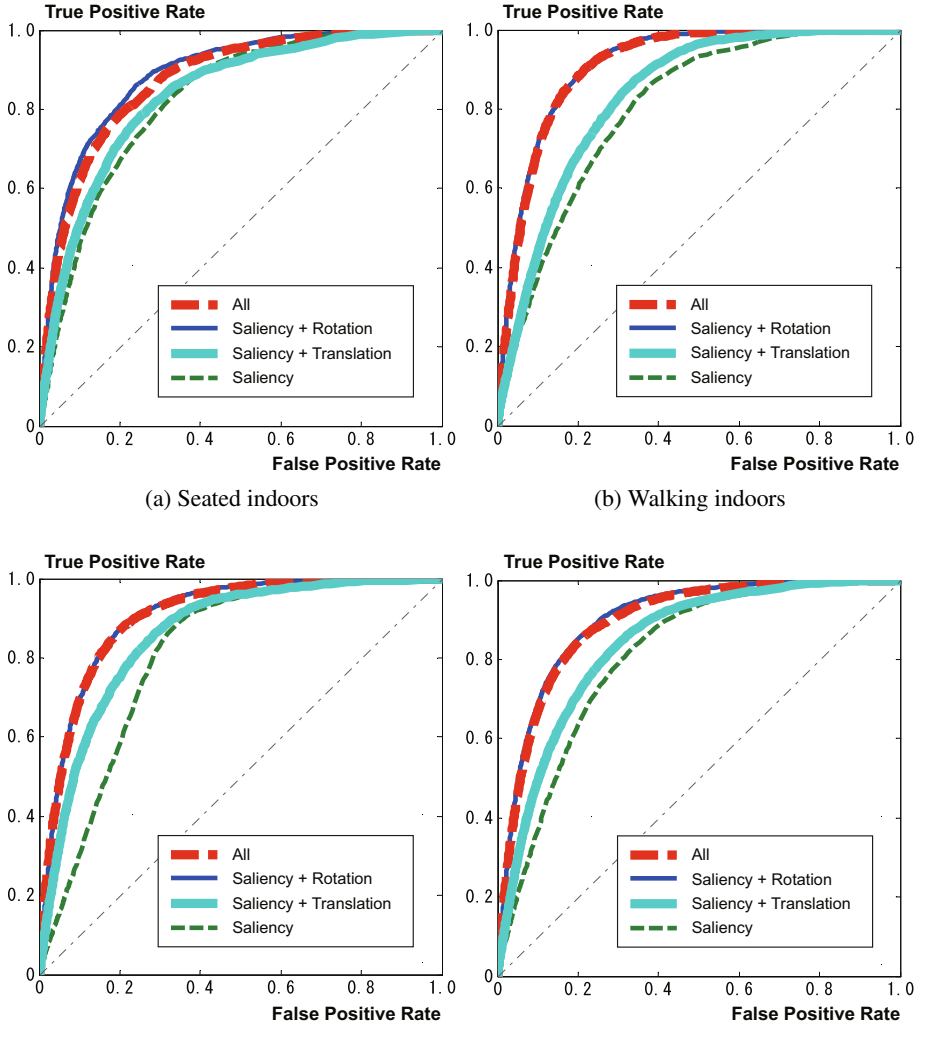
Fig. 6. (a) Mobile gaze tracker employed in experiments. Scene camera is installed and it captures egocentric video of subject at 30 [Hz]. Two eye-cameras and two infrared light sources can record ground-truth gaze points on egocentric video at 240 [Hz]. (b) Example frame of egocentric video. Horizontal field of view of scene camera was 121°, and resolution of egocentric video was 640×480 [pixels].

Egocentric videos and gaze points of five test subjects were recorded under three different settings in which the subjects were: seated indoors, walking indoors, and walking outdoors. Free head movements were allowed in all the settings. Fig. 6(b) shows some examples of the recorded scenes. After rejecting frames with unreliable gaze data caused by actions such as blinking and fast eye movements, the same number of $8,000$ gaze points was selected in each of the 5×3 datasets we used for evaluation.

3.2 Results

To assess how accurately the attention maps predicted a persons' visual attention, we analyzed the correspondence between the maps and the ground-truth gaze points. Fig. 7 shows the ROC curves of the attention maps generated by our framework that were drawn by sweeping the threshold value across all maps. The vertical axis indicates true positive rates, i.e., the rates of gaze points that have higher values than the threshold in the corresponding maps. The horizontal axis indicates false positive rates, i.e., rates of map regions without gaze points that have higher values than the threshold. Therefore, this indicates that the maps can predict gaze points more accurately if the curve approaches the top-left corner.

The area under the curve (AUC) values of the ROC curves are listed in Table 1, where results using a simple centering bias map (centering) have also been listed in addition to the maps (saliency, rotation, and translation) discussed above. It can be seen from these results that our proposed framework can predict actual gaze points more accurately than the standard visual saliency maps and the centering bias maps in egocentric videos. The combination of the visual saliency map and the rotation-based attention map achieved the highest AUC, and thus the highest accuracy.



(a) Seated indoors (b) Walking indoors

(c) Walking outdoors (d) All

Fig. 7. ROC curves of attention maps. Curves were drawn by sweeping threshold value across all maps in four datasets ((a) seated indoors, (b) walking indoorsC(c) walking outdoors, and (d) all combined). Vertical axis indicates true positive rates, i.e., rates of gaze points that have higher value than threshold in corresponding maps. Horizontal axis indicates false positive rates, i.e., rates of map regions without gaze points that have higher value than threshold.

Table 1. Prediction accuracy of attention maps. Each row lists area under curve (AUC) values of ROC curves using bottom-up visual saliency maps (saliency), rotation-based attention maps (rotation), translation-based attention maps (translation), centering bias maps (centering) and their combinations.

Method	AUC
Proposed (saliency + rotation)	**0.900**
Proposed (saliency + translation)	0.841
Proposed (saliency + rotation + translation)	0.893
Saliency	0.809
Rotation	0.892
Centering	0.884
Saliency + centering	0.890

4 Conclusion

We proposed a framework for computing human visual attention maps based on bottom-up visual saliency and egomotion. Rotation-based and translation-based attention maps were generated only using egocentric videos without requiring additional sensors. The effect of using egomotion-based maps was quantitatively evaluated using real egocentric videos, and we demonstrated that the combination of visual saliency maps and rotation-based attention maps could achieve the most accurate predictions of human attention.

Attention prediction using our framework can be done just by using egocentric videos. This has widespread possibilities for applications including casual gaze trackers and attention-based life-log systems. More sophisticated mechanisms for human egocentric visual perception will be investigated in future work to achieve more accurate prediction of visual attention.

References

1. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32(4), 693–708 (2010)
2. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Advances in Neural Information Processing Systems (NIPS), vol. 20, pp. 241–248 (2007)
3. Costa, L.: Visual saliency and atention as random walks on complex networks. ArXiv Physics e-prints, arXiv:physics/0603025, pp. 1–6 (2006)
4. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981)
5. 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(2:6), 1–17 (2008)

6. Fukuchi, M., Tsuchiya, N., Koch, C.: The focus of expansion in optical flow fields acts as a strong cue for visual attention. Journal of Vision 9(8), 137a (2009)
7. Hansen, D., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32(3), 478–500 (2010)
8. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems (NIPS), vol. 19, pp. 545–552 (2006)
9. Hartley, R.: In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 19(6), 580–593 (1997)
10. Hillaire, S., Lécuyer, A., Breton, G., Corte, T.R.: Gaze behavior and visual attention model when turning in virtual environments. In: Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology, VRST 2009, pp. 43–50. ACM, New York (2009)
11. Hillaire, S., Lécuyer, A., Regia-Corte, T., Cozot, R., Royan, J., Breton, G.: A real-time visual attention model for predicting gaze point during first-person exploration of virtual environments. In: Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology, VRST 2010, pp. 191–198. ACM, New York (2010)
12. Itti, L.: Quantitative modeling of perceptual salience at human eye position. Visual Cognition 14(4), 959–984 (2006)
13. Itti, L., Baldi, P.F.: Bayesian surprise attracts human attention. In: Advances in Neural Information Processing Systems, NIPS 2005, vol. 19, pp. 547–554 (2006)
14. Itti, L., Dhavale, N., Pighin, F., et al.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: SPIE 48th Annual International Symposiumon Optical Science and Technology, vol. 5200, pp. 64–78 (2003)
15. 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 (PAMI) 20(11), 1254–1259 (1998)
16. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV), pp. 2106–2113. IEEE (2009)
17. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
18. Ma, Y., Hua, X., Lu, L., Zhang, H.: A generic framework of user attention model and its application in video summarization. IEEE Transactions on Multimedia 7(5), 907–919 (2005)
19. nac Image Technology Inc.: Emr-9, `http://www.nacinc.com/products/Eye-Tracking-Products/EMR-9/`
20. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research 42(1), 107–123 (2002)
21. Qiu, X., Jiang, S., Liu, H., Huang, Q., Cao, L.: Spatial-temporal attention analysis for home video. In: IEEE International Conference on Multimedia and Expo (ICME 2008), pp. 1517–1520 (2008)
22. Shi, J., Tomasi, C.: Good features to track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593–600. IEEE (1994)
23. Tomasi, C., Kanade, T.: Detection and tracking of point features. Carnegie Mellon University Technical Report CMU-CS-91-132, pp. 1–22 (1991)
24. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)
25. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring visual saliency by site entropy rate. In: Computer Vision and Pattern Recognition (CVPR), pp. 2368–2375. IEEE (2010)
26. Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., Hiraki, K.: Can saliency map models predict human egocentric visual attention? In: Proc. International Workshop on Gaze Sensing and Interactions (2010)