Gaze Probing: Event-Based Estimation of Objects Being Focused On

Ryo Yonetani Hiroaki Kawashima Takatsugu Hirayama Takashi Matsuyama
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 6068501, Japan
yonetani@vision.kuee.kyoto-u.ac.jp {*kawashima, hirayama, tm*}*@i.kyoto-u.ac.jp*

***Abstract*—We propose a novel method to estimate the object that a user is focusing on by using the synchronization between the movements of objects and a user's eyes as a cue. We first design an event as a characteristic motion pattern, and we then embed it within the movement of each object. Since the user's ocular reactions to these events are easily detected using a passive camera-based eye tracker, we can successfully estimate the object that the user is focusing on as the one whose movement is most synchronized with the user's eye reaction. Experimental results obtained from the application of this system to dynamic content (consisting of scrolling images) demonstrate the effectiveness of the proposed method over existing methods.**

***Keywords*- event-based gaze estimation, synchronization, dynamic content, eye movement**

I. Introduction

Knowing which object a user is focusing on upon a screen, as he/she selects items on a large list of items, is crucial to understand the user's interests. In this paper, we propose a novel method for estimating the object that a user is focusing on when viewing dynamic content, which consists of scrolling images and text, on a large digital display system (e.g., digital signage).

The majority of existing methods perform a direct comparison between the positions of objects on a screen and the user's eye-gaze points. Identifying an eye-gaze point requires a tradeoff between accuracy and the user's freedom of movement. Hence, with some constraints to the user's head pose and position, we can obtain more accurate results by using the pupil center corneal reflection (PCCR) technique [1]–[4]. However, in a real environment, a user will maintain freedom of motion when focusing on objects. As such, a reactive gaze tracking technique that involves the use of a camera is often more appropriate [5]–[8]. These reactive measurements allow user's head motions in exchange for a large margin of error. In the end, the estimation of objects that a user focuses on has limitations in terms of improving the accuracy.

In this paper, we propose an event-based method, which we refer to as *gaze probing*, to estimate the target objects of a user's gaze by using a designed dynamic content and exploiting the synchronization of motion between objects and the user's eyes as a cue. We first design an *event* as a characteristic pattern of motion, which briefly appears in the movement of each object. In other words, we embed the event in the constituent objects of a dynamic content. We then use the event to infer the target of a user's gaze. If we can successfully design events that are distinct enough to be distinguished from measurement errors, the eye reactions can be detected using data obtained from measurements of gaze direction. We then evaluate the synchronization between eye reactions and events embedded in the movements of objects on a screen; the evaluation is based on the temporal distance between the starting point of each event and that of the detection of eye reaction.

The primary feature of this method is that we do not directly refer to measured gaze data, which sometimes contain significant errors caused by inaccurate measurement systems. We exploit the temporal relationships among events, which represent patterns of motion of the objects and the user's eyes at a more abstract level.

II. Event Design

The accurate detection of the timing of eye motion in response to embedded events is highly dependent upon the design of an event motion pattern. In this section, we therefore discuss various requirements for the design of appropriate event motion patterns in order to ensure that the eye movement reflects the event pattern and that this motion can be distinguished from the measurement errors.

First, we show requirements for the reflection of event motion patterns in gaze data as follows:

A. Short time pattern

As a user changes the focus of his/her gaze frequently, the event should be reflected in gaze data even when an object is focused on over a short time interval. Hence, the event must be a motion pattern that occurs in an interval that is shorter than the time for which the user focuses on an object.

B. Small latency of eye movements

To utilize temporal relations between the starting time of an event and that of a reaction as determined by gaze data, the reaction to the event should occur without significant delay. The event must be a motion pattern that is simple enough for users to predict and must occur at an appropriate speed for them to follow it without difficulty.

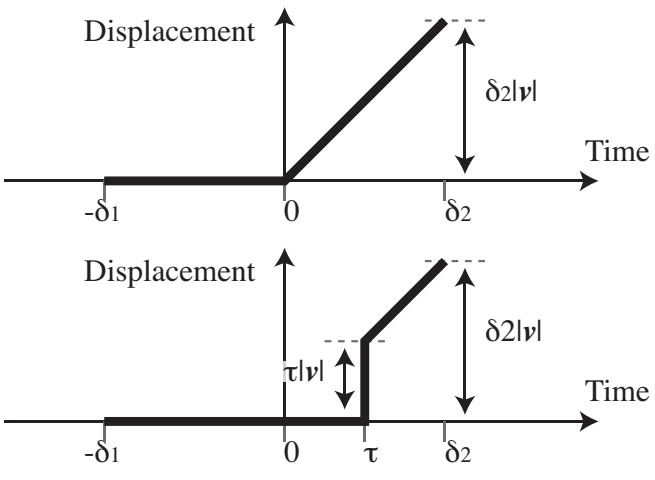

Figure 1. Above: designed motion pattern $\boldsymbol{e}(t)$ around the starting time of an event. Below: eye movement template $\boldsymbol{e}'(t)$ for detecting reactions.

Second, we show requirements for accurately detecting the time of eye reaction as follows:

C. Distinguishable from measurement errors

The spatial range of motion patterns of the event must be large enough so that the eye reaction can be distinguished from measurement errors that arise during gaze tracking.

D. Distinguishable from endogenous eye movements

Gaze data often contains not only reactions but a user's endogenous eye movements, which can be complex patterns caused by the user's examination of each object. The spatial range of motion patterns of the event should therefore be larger than the object size. Endogenous eye movements also occur when the user changes a target of his/her gaze. Hence, the motion direction of the event and the orientation of constituent objects should be as orthogonal as possible.

In this paper, we use the "onset of object scroll" as a basic motion pattern for the event that fulfills the aforementioned requirements (see Figure 1). Around the starting time of the event, objects move on the basis of the following equation:

$$\boldsymbol{e}(t) = \begin{cases} \boldsymbol{0} & (-\delta_1 \leq t < 0) \\ t\boldsymbol{v} & (0 \leq t \leq \delta_2), \end{cases} \tag{1}$$

where the starting time is at $t = 0$. Objects in which this event is embedded keep on stopping within the time interval $[-\delta_1, 0]$ and begin scrolling from time 0 with a constant velocity $\boldsymbol{v}$.

Due to requirement II-A, the time interval $(\delta_1+\delta_2)$ allotted for the movement must be shorter than the period of a user's visual fixation upon any given object, where the fixation time depends upon an object's type of media, complexity, size, and shape. In the experiments, we thus determine the fixation time by considering the actual objects in use. The maximum speed at which a human eye is capable of tracking an object smoothly is around $40°$/s. Therefore, due to requirement II-B, $|\boldsymbol{v}|$ must be strictly less than this speed. Due to requirements II-C and II-D, the scrolling range $\delta_2 |\boldsymbol{v}|$ over time interval δ_2 must be strictly greater than the gaze tracking error and endogenous eye movements investigating objects. This range can be determined by considering the precision of a gaze tracking system when working with objects that we actually use. In addition, the directions of the object motion patterns are set to be horizontal while the orientation of the objects is set to be vertical.

We embed this event into the movement of objects $\{O_n | n = 1, \ldots, N\}$. Let the centroid of O_n be $\boldsymbol{x}_n \in \mathrm{R}^2$, and let the ith starting time of the event on O_n be $t_{(n,i)}$. Then, the movement of O_n around $t_{(n,i)}$ ($[t_{(n,i)} - \delta_1, t_{(n,i)} + \delta_2]$) is determined as follows:

$$\boldsymbol{x}_n(t) = \boldsymbol{e}(t - t_{(n,i)}) + \boldsymbol{x}_n(t_{(n,i)}). \tag{2}$$

For different objects $O_m, O_n (m \neq n)$, we set assume that the starts of the event are defined at different times as $\left|t_{(m,i)} - t_{(n,i)}\right| > \varepsilon$. The time interval of each start, ε, is larger than the latency of eye movement. Thus, we can estimate the target object of a user's gaze if a reaction is detected in the gaze data. Note that object movements other than events are not specifically defined here; these movements are designed with lower salience than the motion patterns of the event.

III. Estimation of Objects Being Focused On

We assume that a user focuses on the object in which the motion pattern of event $\boldsymbol{e}(t)$ is embedded. A sequence of eye-gaze points is captured as gaze data, $\boldsymbol{X}(t)$, by a gaze tracking system. Then, we can detect users' reactions to the event using a template $\boldsymbol{e}'(t)$. This template depicts the eye movement pattern that appears in $\boldsymbol{X}(t)$ when a user tracks $\boldsymbol{e}(t)$. Humans tend to track moving objects as follows: they generally identify and track an object in the central fovea, exhibiting a catch-up saccade after latency τ of about 0.15 s, and continue tracking the object with decreasing retinal slips [9]. Taking this point into consideration, we define the eye movement pattern $\boldsymbol{e}'(t)$ in Eq. (3) (See Figure 1).

$$\boldsymbol{e}'(t) = \begin{cases} 0 & (-\delta_1 \leq t < \tau) \\ t\boldsymbol{v} & (\tau \leq t \leq \delta_2). \end{cases} \tag{3}$$

We first detect the user's reaction time T using the correlation based template matching in $\boldsymbol{X}(t)$, and identify the events that occur around T. We then calculate the temporal distance between the starting time of each event $t_{(n,i)}$, and T. If several events are detected, we compare the temporal distances for each event. We can estimate the target of a user's gaze as the object whose movement is most

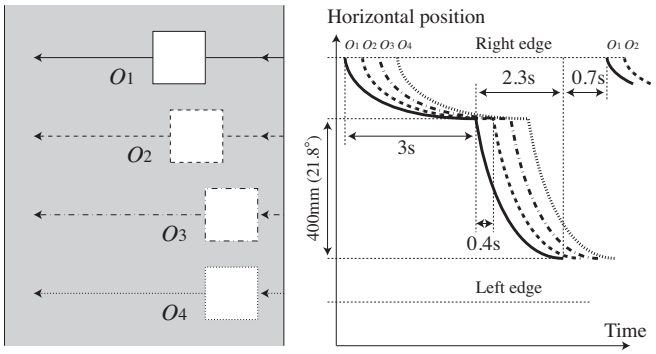

Figure 2. D-1 Scrolling design

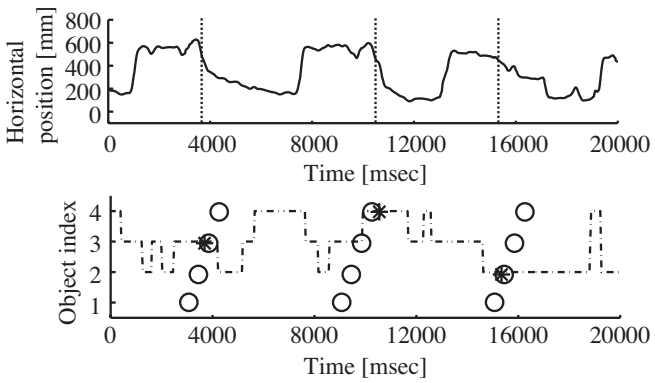

Figure 3. Above: gaze data (solid line) and reactions (dot line). Below: embedded events (o), reactions (*) and the verified object (dashed line).

synchronized with the user's reaction. For example, when a user gazes at the pth event in O_k and a reaction is detected at time T, the following relation can be derived:

$$(k, p) = \arg\min_{n,i} \left| T - t_{(n,i)} \right|. \quad (4)$$

IV. Experiments

A. Experimental Setup

To evaluate the accuracy of the proposed method, we designed catalog content D-1 and D-2 depicting photo images of cellular phones (150 mm $\times$ 150 mm) with a small caption in Japanese (around 50 characters) on a screen (1106 mm in height and 633 mm in width). Six subjects were asked to choose their favorite object from a set of four objects displayed on the screen.

With regard to D-1, as shown in Figure 2, each object appeared from the right edge of the screen, paused for a short period of time, and then scrolled to the left. As an object approached the left of the screen, it slowed down in a smooth motion, before stopping at the screen edge and finally disappearing. Following this, the object appeared at the right edge of the screen once again, and moved in the same manner. There were three photo images displayed for each object; these images updated the contents of the object at each new appearance of the object from the right of the screen. Events in this case were defined as the scrolling movement from stopping at the right edge. The time interval between events for a given object was set to 6 s, and the period between the events of two objects was set to 0.4 s.

With regard to D-2, each object swung from the left edge to right edge. First, each object scrolled to the left. As an object approached the left edge of the screen, the object smoothly slowed until it stopped at the left edge for a short period of time. The object then scrolled to the right in a similar manner. Events in this case were defined as the scrolling movement from stopping at each edge. The time interval between events for a given object was set to 3 s, and the period between events for two different objects was set to 0.4 s.

Table I
Estimation accuracies

	M_{pos}[%]	M_{squ}[%]	Proposed [%]
D-1	41.9	51.4	**76.8**
D-2	54.5	63.3	**68.4**

A simple gaze tracking system using a single camera was employed to obtain eye-gaze points. A camera (UXGA, 30 fps, 8 bit gray scale) was situated below the screen. The subjects were allowed to change the pose and position of their head while the camera recorded their face. First, we detected the facial region and extracted facial features from the captured image sequence, fitted the features to a 3D face shape model that was calibrated with stereo cameras during a preliminary experiment, and detected the iris regions. We then estimated the 3D positions of the eyeball centers and that of the iris centers. Finally, we obtained a sequence of eye-gaze points $\boldsymbol{X}(t)$ as the intersections of the screen and a straight line running through both the eyeball and iris center. The gaze tracking accuracy was 62 mm in average (3.6°) along the horizontal direction, and 94 mm (5.4°) along the vertical direction. In addition to the above gaze tracking system, in order to get the ground truth of verify the estimated object, we used another gaze tracking system that is based on the PCCR technique (60 Hz with an approximate allowed range of head motion within 400 $\times$ 220 $\times$ 300 mm). The gaze tracking accuracy of the PCCR system was, on average, around 27 mm (1.6°).

We employed the eye movement template $\boldsymbol{e}'(t)$ (see Eq. (3)). τ was set to 0.15 s to compensate for a latency, δ_1 was set to 0.6 s taking into account the object complexity, and δ_2 was set to 0.64 s taking into account the gaze tracking accuracy and the object size (in this setting, $\delta_2 |\boldsymbol{v}|$ was set to 300 mm).

B. Results

As shown in Figure 3, the events caused reactions that were easily distinguishable from endogenous eye move-

ments. 56 reactions were detected for D-1 (93.3% reactions detected) and 98 for D-2 (90.7% detected) during the course of the 360 s experiment. The system's accuracy is calculated as the ratio of the number of correctly estimated reactions to the total number of detected reactions. We use two evaluation methods:

- $\mathrm{M_{pos}}$ estimation using the distance between eye-gaze points and object positions; this involves the calculation of accuracy as the ratio of the number of correctly estimated frames to the total frames,
- $\mathrm{M_{squ}}$ estimation using the distance and the similarity between the temporal patterns of objects' motion and those of eye movement based on the squared value of the distance of the patterns; this involves the calculation of the accuracy in the same manner as that in the proposed method.

The accuracies calculated using each method are shown in Table I. Measurement error in the gaze data $\boldsymbol{X}(t)$ had little effect on the proposed method so that it shows greater accuracy than previously used evaluation methods. Other evaluation methods were found to often fail in their estimation, since the distance between the objects was smaller than the errors in $\boldsymbol{X}(t)$.

In contrast to $\mathrm{M_{pos}}$, the proposed method is an event-based, and therefore intermittent, estimation method. We interpolate the estimation result for the proposed method as follows: we first detect changes in the target of the user's gaze. We then define fixation intervals between successive times of change. The estimation results can be applied to the given fixation intervals if the reactions take place within those intervals (see Figure 4). Using the dynamic content within the experiment, we detected changes in the object focus of a subject's gaze using observed accelerations of eye movement along a vertical direction. The estimation accuracies within the applied intervals were 61.6% (45.4% applied) using content D-1 and 63.4% (74.4% applied) using D-2. As the error associated with vertical gaze tracking approached the value of the distance between objects, the detection of changes in user's objects of focus became unstable and thus the estimation accuracy decreased. However, the estimation accuracy was still greater than that of the $\mathrm{M_{pos}}$ evaluation method.

V. Conclusions

We proposed an event-based method known as gaze probing to estimate the object that is the focus of a user's gaze. Experimental results revealed that our method, which exploits the temporal relationship among embedded events and the user's ocular reactions, is more accurate than existing position-based methods. Future work will seek to apply the proposed method to interactive display systems that move objects to attract a user's attention, estimate his/her interests based on whether he/she focuses on those objects, and recommend items to suit to the user's interests.

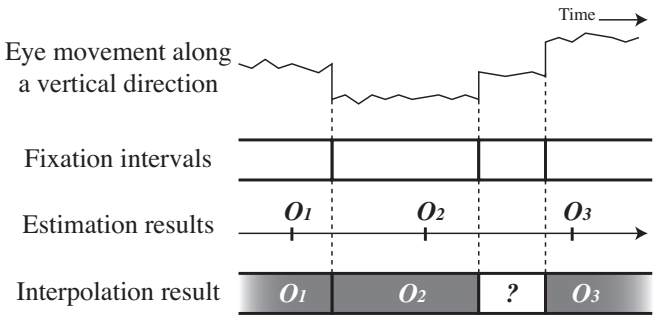

Figure 4. Interpolation of estimation results using observed accelerations of eye movement along a vertical direction.

Acknowledgment

This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 18049046.

References

[1] J. Chen, Y. Tong, W. Gray, and Q. Ji, "A robust 3d eye gaze tracking system using noise reduction," *Proc. of the symposium on Eye tracking research & applications*, pp. 189–196, 2008.

[2] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," *Proc. of the symposium on Eye tracking research & applications*, pp. 87–94, 2006.

[3] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 4–24, 2005.

[4] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 918–923, 2005.

[5] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, no. 2, pp. 451–458, 2003.

[6] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models," Robotics Institute, Tech. Rep., 2004.

[7] J.-G. Wang, E. Sung, and R. Venkateswarlu, "Estimating the eye gaze from one eye," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 83–103, 2005.

[8] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe, "Remote and head-motion-free gaze tracking for real environments with automated head-eye model calibrations," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, vol. 0, pp. 1–6, 2008.

[9] C. Rashbass, "The relationship between saccadic and smooth tracking eye movements," *The Journal of Physiology*, vol. 159, no. 2, pp. 326–338, 1961.