Vision Research 117 (2015) 59–66

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Saliency-based gaze prediction based on head direction

Ryoichi Nakashima [*],[1], Yu Fang, Yasuhiro Hatori, Akinori Hiratani, Kazumichi Matsumiya, Ichiro Kuriki, Satoshi Shioiri

Tohoku University, and CREST, Japan Science and Technology Agency, Japan

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form 1 October 2015
Accepted 5 October 2015
Available online 8 November 2015

Keywords:
Gaze prediction
Head direction
Eye–head coordination
Saliency map

ABSTRACT

Despite decades of attempts to create a model for predicting gaze locations by using saliency maps, a highly accurate gaze prediction model for general conditions has yet to be devised. In this study, we propose a gaze prediction method based on head direction that can improve the accuracy of any model. We used a probability distribution of eye position based on head direction (static eye–head coordination) and added this information to a model of saliency-based visual attention. Using empirical data on eye and head directions while observers were viewing natural scenes, we estimated a probability distribution of eye position. We then combined the relationship between eye position and head direction with visual saliency to predict gaze locations. The model showed that information on head direction improved the prediction accuracy. Further, there was no difference in the gaze prediction accuracy between the two models using information on head direction with and without eye–head coordination. Therefore, information on head direction is useful for predicting gaze location when it is available. Furthermore, this gaze prediction model can be applied relatively easily to many daily situations such as during walking.

1. Introduction

Humans cannot simultaneously process the vast amount of visual information they receive in daily life, so they must select which incoming visual information to process in detail. This selection process typically consists of saccadic gaze shifts that fixate on different regions of a visual scene projected onto the fovea. To predict gaze location, especially gaze with attention focused at a location, a number of models using saliency maps, which topographically represent the visual saliency of a given scene, have been proposed (Itti, Koch, & Niebur, 1998). Saliency maps are based on the bottom-up architecture of visual attention proposed by Koch and Ullman (1985), which involve the hypothesis that the most salient locations in a visual scene tend to attract attention. Visual saliency is calculated by integrating visual features of a scene, such as color, luminance, and orientation, often with consideration variety of visual functions, like retinal inhomogeneity (Kubota et al., 2012) and the canceling out of self-motion (Hiratani, Nakashima, Matsumiya, Kuriki, & Shioiri, 2013). However, the accuracy of gaze prediction using visual saliency alone is limited because it is based on bottom-up factors such as visual features, and does not account for the influence of top-down factors such as the intention of the observer (e.g., Henderson, Brockmole, Castelhano, & Mack, 2007; Peters & Itti, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006; see also Kimura, Yonetani, & Hirayama, 2013).

Models that account for top-down factors provide better gaze prediction. One representative method employs a machine learning technique to identify additional information for possible locations of gaze location from empirical data (e.g., Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Torralba et al., 2006). Models utilizing learning techniques are effective when the task and scenes are known, making learning possible beforehand, for example, when searching for people in outdoor scenes.

In this study, to improve the accuracy of gaze (and attention) prediction, we propose a method utilizing the natural human behavior of head direction. Eye and head movements are typically coordinated, as observed during simple gaze shifts to targets present in the periphery (e.g., Cecala & Freedman, 2008; Freedman, 2008; Freedman & Sparks, 2000; Fuller, 1992; Oommen, Smith, & Stahl, 2004; Stahl, 1999; Thumser, Oommen, Kofman, & Stahl, 2008; Zangemeister, Jones, & Stark, 1981). In these studies, eye–head coordination was as follows when gaze shifts were sufficiently large. When an observer shifted gaze to the left (right), the head moved to the left (right) and eye movement was to the left (right) relative to the head (Stahl, 1999). This indicates that head direction biases eye position. Additionally, we previously found that complex tasks such as visual search tasks involve coordinated movements of the eyes and head (Fang, Nakashima,

[*] Corresponding author at: Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan.
E-mail address: one.island.one@gmail.com (R. Nakashima).

[1] Ryoichi Nakashima is now at RIKEN as a postdoctral researcher.

http://dx.doi.org/10.1016/j.visres.2015.10.001
0042-6989/© 2015 Elsevier Ltd. All rights reserved.

Matsumiya, Kuriki, & Shioiri, 2015); therefore, we expected that head direction would be useful for predicting gaze location during general viewing conditions. This expectation was also based on a report that visual processing is modulated by head direction (Nakashima & Shioiri, 2014, 2015), which suggests that a specific eye–head relationship can influence visual processing. In the method proposed in this study, we weight saliencies, which attract attention, in the map according to a probability distribution of eye position that is estimated based on head direction. Head direction is not estimated in this method and thus needs to be measured using a device such as a monitoring camera.

2. Experiment

We conducted an experiment to investigate the relationship between head direction and eye position during the viewing of large images of natural scenes to formulate eye–head coordination for gaze prediction. This is in contrast to previous studies on eye–head coordination, most of which analyzed single-step gaze shifts (e.g., Cecala & Freedman, 2008; Freedman, 2008; Freedman & Sparks, 2000; Fuller, 1992; Oommen et al., 2004; Stahl, 1999; Thumser et al., 2008; Zangemeister et al., 1981), which are inappropriate for predicting the large and continuous gaze shifts that occur in everyday life. We investigated both horizontal and vertical components of eye and head movements so that our results could be applied in two dimensions. Although some studies regarding vertical eye–head movement coordination have been conducted (Freedman, 2005; Goossens & Van Opstal, 1997; Tweed, Glenn, & Vilis, 1995), no systematic comparisons in relation to continuous gaze shifts have been made; therefore, no adequate data were available for the purposes of this study.

2.1. Method

2.1.1. Observers

This experiment was conducted during an outreach activity at the National Museum of Emerging Science and Innovation in Tokyo, Japan. Study participants comprised 228 museum visitors (92 females; mean ± *SD* age: 21.2 ± 15.2 years). All the participants had normal or corrected-to-normal vision. This experiment was approved by the institutional review board of Tohoku University, and written informed consent was obtained from all observers. This experiment was conducted in accordance with the Declaration of Helsinki in the treatment of the observers.

2.1.2. Apparatus

Visual stimuli were generated with a computer using the Psychophysics Toolbox for MATLAB (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997), and displayed on a 100-inch screen using a short throw projector (NP-U310WJD; NEC, Japan). FASTRAK (60 Hz; Polhemus, USA), an electromagnetic motion tracking system, was used to track the direction (azimuth, elevation) of one small sensor, which was secured to the head of the observer to record head direction. Eye movements and positions were recorded at a sampling frequency of 60 Hz by an eye tracker (EMR-9, NAC, Japan) equipped with two cameras for recording the positions of the eyes and a scene camera with a 62° field-of-view. A computer controlled the experimental sessions, including temporal synchronization among display presentations, as well as head direction and eye position measurements.

2.1.3. Stimuli

A total of 30 natural scenes (6 indoor and 24 outdoor) containing numerous objects (see Fig. 1a and b) were prepared as stimuli and projected onto a large screen. The size of each image was designed to be 57° × 44° from a viewing distance of 125 cm. The images were numbered from 1 to 30 in advance, and divided into 10 stimuli sets, each of which included 3 images (Set 1: images 1–3, Set 2: images 4–6, Set 3: images 7–9, etc.).

2.1.4. Procedure

The experiment was performed in an illuminated area, but without direct illumination on the screen. An observer sat on a chair in front of the screen (Fig. 1c). The viewing distance was set at 125 cm when the observer oriented their head straight toward the screen. It should be noted that the viewing distance varied to some extent throughout the session as the observer moved their head to look at different regions of the screen. The sensor and eye tracker were fitted on the observer, and calibration was performed before the experiment.

The observer was instructed to view each image displayed on the screen for 5 s and to memorize it for a later task. After memorizing one image, the observer took part in a change blindness experiment (cf. Nakashima & Yokosawa, 2012; Rensink, O'Regan, & Clark, 1997). One part of the image was changed to make the second image, and the observer was then asked to detect the difference between the original and changed images while the two were alternatively presented for 250 ms, followed by blank gray screen for 250 ms. Each observer viewed a set of three images selected from the 30 images in advance (i.e., one of the stimuli sets), and eye position and head direction were recorded during the 5 s for memorizing each image. We did not design the experiment to analyze data during the change blindness experiment because we found from a pilot observation of eye tracking data that the accuracy was low. Observers were often excited and made head movements and facial expressions that caused eye movement recordings to be unstable.

2.2. Results and discussion

We obtained eye position relative to head direction, and head direction relative to the space (i.e., body direction). To ascertain directional differences, we analyzed the horizontal and vertical components of the eye position and head direction data separately. Fig. 2 shows the horizontal and vertical distributions of head direction. The head was oriented within ±12° in the horizontal and vertical directions in most cases (plus means "right" in the horizontal data and "up" in the vertical data). Furthermore, about the half of the head direction was concentrated around the center (within ±3°), 43.1% in the horizontal dimension, and 56.8% in the vertical dimension, perhaps due to the tendency of the observers to maintain a natural posture without much tension during the task.

We analyzed eye position relative to head direction, which was the output of the eye tracker. The results showed that more than 95% of eye fixations were recorded when the head direction was within ±12°; therefore, we analyzed the relationship between eye position and head direction within this range. For the initial proposal regarding the use of head direction for estimating gaze positions with an attention model, the purpose of this experiment was to investigate the relationship between the distributions of head direction and eye position when both the eyes and head were stationary (i.e., static eye–head coordination). The head was considered stationary when the velocity was less than 3°/s. Next, to determine the onset and end of saccades for defining the fixation position (eye position), we calculated the velocity and acceleration of gaze movements. Saccade onset was defined as the time when both gaze velocity and acceleration exceeded a velocity threshold of 75°/s and an acceleration threshold of $200°/s^2$. Saccade end was defined as the time when both the velocity and acceleration fell below their respective thresholds (see Fang, Nakashima, et al., 2015). Fixation positions were defined by averaging over