The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Prediction of Human Eye Movements in Facial Discrimination Tasks

S. Nishida, T. Shibata and K. Ikeda
Graduate School of Information Science Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192
{satoshi-n,tom,kazushi}@is.naist.jp

Abstract

Under natural viewing conditions, human observers selectively allocate their attention to subsets of the visual input. Since overt allocation of attention appears as eye-movements, the mechanism of selective attention can be uncovered through computational studies of eye-movement prediction. Since top-down attentional control in a task is expected to modulate eye-movements signi cantly, the models that take a bottom-up approach based on low-level local properties are not expected to suffice for prediction. In this study, we introduce two representative models, apply them to a facial discrimination task with morphed face images, and evaluate their performance by comparing them with the human eye-movement data. The result shows that they cannot predict well the eye-movements in this task.

1 Introduction

Surrounded by complicated visual information, human visual processing selectively allocates limited computing resources such as spatial/feature/object attention. Eye-movements achieve overt allocation of attention by shifting the fovea to the interesting region of a visual scene. Therefore, eye movements are representative of overt attention at the behavioral level, and the mechanism of selective attention can be uncovered through computational models for eye-movement prediction.

Eye-movement controllers are roughly divided into the bottom-up control caused by external factors and top-down control caused by internal factors. While the bottom-up control is based on a static mechanism based on low-level image features such as color, orientation, intensity and so on, the top-down control is based on a dynamic mechanism that depends on semantics, context or task-related factors treated by the high-level cognitive function in the brain.

Recently, a couple of biological models for predicting eye-movements were proposed. One of these models is the saliency-based model[3]. Alternative models utilize Shannon's information theory. These models are designed to explain eye-movements from the viewpoint of minimizing uncertainty in the visual information. Renninger's model[6] is representative of these. In this study, we focus on the saliency-based model and Renninger's model. All of these models are based on bottom-up control.

In this paper, we introduce these two models, and apply them to a discrimination task as well as a free-viewing task of faces to illustrate the possibility of eye-movement prediction in these tasks. Finally, we discuss some characteristics of human eye-movements speci c to facial recognition from the viewpoint of feature selectivity.

2 Methods

2.1 Behavioral Task

Our tasks consisted of a discrimination task and a free-viewing task. One of the purposes of our tasks is verifying that the bottom-up models cannot predict well human eye-movements in facial recognition. The other is analyzing how the goal of discrimination modulates feature selective strategy in the context of facial recognition through comparing the results of these two tasks.

Figure 1 shows a set of morphing images for use in these tasks. To make the set, we generated sixteen morphing images that had continuous change of the mixing ratio of two faces, and extracted eight images that were closer to the half ratio. In the same way, four sets (32 images) were created for these tasks. They were based on face images of six men and two women ranging in age from 22 to 27 years old. The pairs were made of same-sex face images. For morphing, each face image was converted to a gray-scale image, and the background was painted black. The

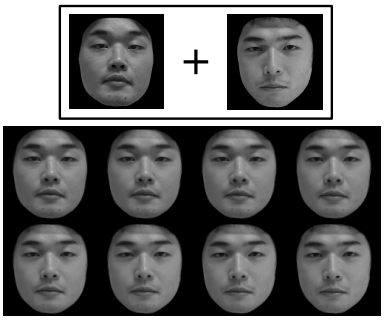

Fig. 1: A set of facial morphing images.

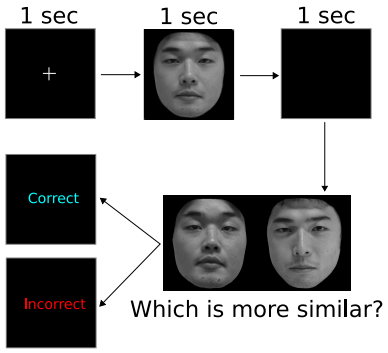

Fig. 2: Procedure of each trial in the facial discrimination task.

morphing images were created by a morphing software, WinMorph[4], whose algorithm is a eld morphing technique[1].

Subjects were seated 73.5 cm from the display screen and were put on an Eyelink II (SR Research) eye-tracking device. The screen had a size of 31 cm 25 cm, a visual angle of 23.8° 19.3°, a resolution of 1024 pixels 768 pixels and a frame-rate of 59.84 Hz. Eye-position data were acquired at 500 Hz and both eyes were tracked. The stimuli were presented by Matlab's Psychophysics and Eyelink toolbox extensions[2].

In the discrimination task, each session consisted of 8 blocks of 16 trials. At the beginning of each block, calibration of the eye-tracker was executed, and at the beginning of each trial, drift correction was executed. Figure 2 sketches the procedure of each trial. On each trial, subjects pressed a button to begin. First, one of the 32 morphing images (12.5° 12.5°) was presented for 1 sec. It was chosen randomly, but all images appeared at the same times in a session. Then, two original face images (12.5° 12.5°) were displayed together. Until subjects pressed a button to choose the face that was more similar to the rst morphing image, the images did not disappear. Finally, feedback was given.

In the free-viewing task, stimuli were presented under neutral viewing condition without an explicit task goal. Each session consisted of 8 blocks of 16 trials. Calibration and drift correction of the eye-tracker were done in the same way as the discrimination task. On each trial, subjects pressed a button to begin, then only the morphing image (12.5° 12.5°) was displayed for 1 sec.

Three male subjects participated in the experiment. Subjects ranged in age from 23 to 43 years old. All subjects had normal eyesight.

2.2 Models for Eye Movement Prediction

Eye-tracking data in the experiment were compared with simulation results of the following models.

2.2.1 Saliency-based Model

This is a model of the visual bottom-up attention mechanism for early visual processing in primates. An input image is decomposed into a set of feature maps, followed by center-surround di erences and normalization of three features (intensity, color and orientation). All feature maps are then combined into a unique topographic saliency map. The winner-take-all network detects the most salient location and directs attention toward it. An inhibition-of-return mechanism transiently suppresses this location in the saliency map, such that attention is autonomously directed to the next most salient image location. This model can plausibly explain eye-movements of the bottom-up control under the context-free viewing condition.

2.2.2 Renninger's Model

This is a model for eye-movement prediction in object discrimination tasks with silhouette images. However, it is also a bottom-up model because the goal of discrimination was not adopted into the evaluation function. One of its characteristics is adoption of foveal and peripheral vision mechanisms. Since discrimination objects are silhouettes, the information needed for the task is the edge orientations. Consequently, the strategy of this model for xation selection is minimization of entropy within edge orientations with respect to variable resolution in the visual eld.

The algorithm of this model is described below. First, edges are decomposed into a collection of

edgelets, each of which has one of eight possible orientations. Each edgelet j is given a local region whose size depends on eccentricity $E_j(\mathbf{F})$ from the current xation point $\mathbf{F}$ (using parameters from the vernier acuity literature[5]). The probabilistic distribution of the orientation x_j of edgelet j is generated by the histogram $\mathbf{h}_j(\mathbf{F})$ regarding orientations of all edgelets within the region.

$$P(x_j|\mathbf{h}_j(\mathbf{F}), E_j(\mathbf{F})) = h_{j,x_j}(\mathbf{F})/Z, \quad (1)$$

where Z is a normalization constant.

Then, a resolution-dependent entropy (RDE) of each pixel i is computed.

$$\mathrm{RDE}_i = \sum_{j \in \text{all edgelet locations within radius } r(E_i(\mathbf{F})) \text{ of } i} H_j, \quad (2)$$

where $r(E_i(\mathbf{F}))$ is the radius of the circular region determined by the eccentricity $E_i(\mathbf{F})$ from the xation point $\mathbf{F}$, and the entropy H_j of edgelet j is computed by

$$H_j = \sum_{z=1}^{8} P(x_j = z) \log P(x_j = z). \quad (3)$$

Thus, an RDE map, which represents the uncertainty of shape knowledge at any point, is generated. The next xation is directed towards the maximum point of the map. The new probabilistic distributions, depending on a new xation point, are integrated with the old ones by the Bayesian rule. The posterior probability can be updated for multiple xations $\mathbf{F}_1$ and $\mathbf{F}_2$ by

$$P(x_j|\mathbf{h}_j(\mathbf{F}_1), \mathbf{h}_j(\mathbf{F}_2), E_j(\mathbf{F}_1), E_j(\mathbf{F}_2)) = h_{j,x_j}(\mathbf{F}_1) h_{j,x_j}(\mathbf{F}_2)/Z', \quad (4)$$

where Z' is a normalization constant.

In addition, this model adopts a human property that saccades to a simple shape or object often landing near the centroid of that object.

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = w \begin{bmatrix} C_x \\ C_y \end{bmatrix} + (1 \quad w) \begin{bmatrix} \hat{f}_x \\ \hat{f}_y \end{bmatrix}, \quad (5)$$

where f is the next xation, $\hat{f}$ is the model-de ned prediction, $\mathbf{C}$ is the centroid, and w is a weight.

3 Results

Three subjects' accuracy rates ranged from 85.9% to 89.5%, and their response times ranged from 0.61 sec to 0.84 sec. Figure 3 shows the average accuracy rate and the average response time of the subjects with respect to distance between two faces. The distance is de ned as the di erence between each mixing ratio of the two source faces of the morphing image, supposing that the distance between the two source faces is 1. The accuracy rate increased and the response time decreased with increasing distance.

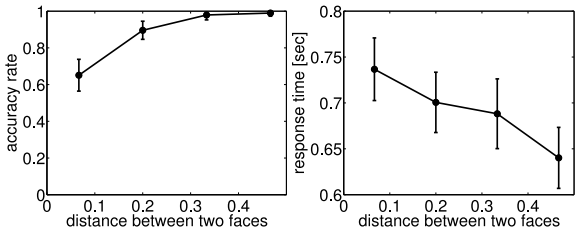

Fig. 3: Accuracy rate (left) and response time (right) with respect to distance between two faces. Error bars indicate 95% con dence intervals.

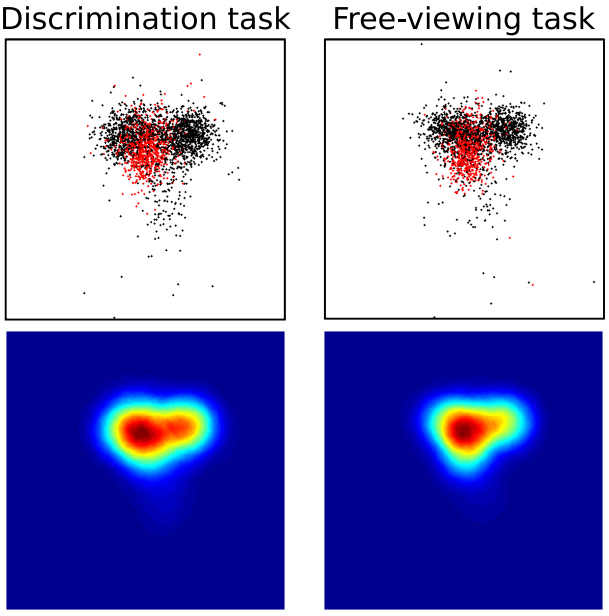

Fig. 4: Fixation distribution (upper) and foveated density (lower).

The upper part of Fig.4 shows the xation distribution. Red points indicate the rst xations. Most of the xations concentrated around the eyes and noses. From the middle points of the eyes, 90% of all xations are distributed within a visual angle of 2.4° and 90% of the rst xations are distributed within a visual angle of 1.9°. The lower part of Fig.4 shows the foveated density generated by the xation distributions with

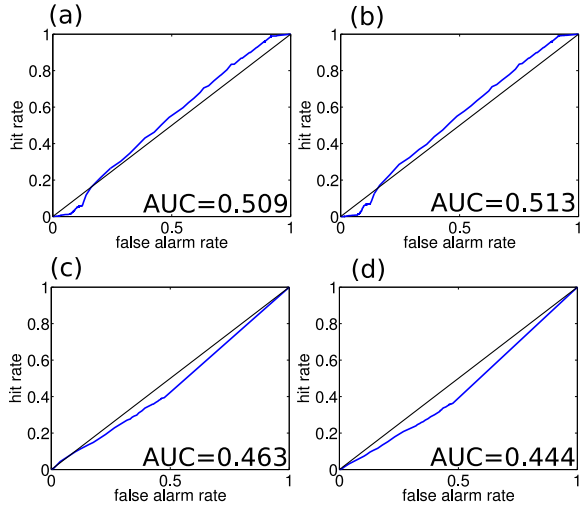

Fig. 5: ROC analysis. Saliency-based model (a,b). Renninger's mode (c,d). Discrimination Task (a,c). Free-viewing Task (b,c)

respect to the visual angle of fovea (2°).

The predictability of the two models discussed above was evaluated with ROC analysis. Figure 5 shows ROC curves and the area-under-the-curve (AUC) of the saliency-based model and Renninger's model (without a centroid bias).

4 Discussion

We found that most xations were concentrated around the eyes and noses under the condition of face recognition. This is also supported by the results that the predictability of Renninger's model with a centroid bias (Eq.5) was higher than one without a centroid bias. Although w ranging from 0.2 to 0.3 made the AUC highest in Renninger's experiment[6], the highest AUC = 0.573 is obtained by $w = 0.5$ in our discrimination task and the highest AUC = 0.565 is obtained by $w = 0.6$ in our free-viewing task. However, this cannot lead to the conclusion that Renninger's model with a centroid bias is compatible with eye-movements in our tasks. Since psychological or physiological plausibility of a centroid bias in our tasks has not been unproved, this is just a heuristic approach.

In comparison between the discrimination task and free-viewing task (Fig.4, Fig.5), we could not see a clear di erence. This result de ed our expectation that the variability of eye-movements would become greater to obtain more features needed by discrimination. There are two possibilities for the conclusion that this result leads to. One is that using peripheral vision is sufficient to obtain facial features. The other is, in contrast, that the human eye-movement strategy during face viewing involuntarily has the goal of discrimination. These possibilities cannot be examined with our experimental paradigm, but need to be con rmed in further work.

We found that it is difficult for the existing bottom-up models to predict eye-movements in our tasks. The eye-movement predictability of the saliency-based model and Renninger's model is not so di erent from one of random strategy. To explain eye-movements during face viewing, we must clarify a speci c feature space of the face, and consider the mechanism from the viewpoint of feature selection.

Acknowledgements

This work was partly supported by Grant-in-Aid for Scienti c Research from Japan Society for the Promotion of Science, No. 18300101.

References

[1] T. Beier and S. Neely. Feature-based image metamorphosis. *ACM SIGGRAPH Computer Graphics*, Vol. 26, No. 2, pp. 35–42, 1992.

[2] F. W. Cornelissen, E. M. Peters, and J. Palmer. The eyelink toolbox: eye tracking with MATLAB and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers*, Vol. 34, No. 4, pp. 613–617, 2002.

[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 1254–1259, 1998.

[4] S. Kumar. Winmorph [computer software]. http://www.debugmode.com/winmorph/, 2002.

[5] D.M. Levi, S.A. Klein, and AP Aitsebaomo. Vernier acuity, crowding and cortical magni cation. *Vision Research*, Vol. 25, No. 7, pp. 963–977, 1985.

[6] LW Renninger, P. Verghese, and J. Coughlan. Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, Vol. 7, No. 3, pp. 1–17, 6 2007.