Studying Human Face Recognition with the Gaze-Contingent Window Technique

Naing Naing Maw (nnmaw@cs.umb.edu)
University of Massachusetts at Boston, Department of Computer Science
100 Morrissey Boulevard, Boston, MA 02125-3393, USA

Marc Pomplun (marc@cs.umb.edu)
University of Massachusetts at Boston, Department of Computer Science
100 Morrissey Boulevard, Boston, MA 02125-3393, USA

Abstract

In eye-movement experiments using gaze-contingent windows, the stimulus display is continuously updated in response to the participant's current gaze position. Usually, a window is centered at the participant's gaze position and follows it wherever the participant looks. Within the window, all stimulus information is visible, while outside of the window at least part of the information is masked. In the present paper, we apply this technique to a face recognition task. By varying the size of the window, we gain insight into face recognition processes in humans and characterize the visual information on which face recognition relies. The results also motivate the use of gaze-contingent windows to study visual perception.

Introduction

Face recognition is a very important function of the human visual system and is fundamental to our complex social behavior. Therefore, it is not surprising that face recognition in humans has been extensively studied. Many studies concluded that face recognition relies more strongly on holistic information than does object recognition in general (see Maurer, Le Grand & Mondloch, 2002, for a review). In other words, ideally, the recognition process uses the entire visual information available from a face.

What makes faces so special in this regard? An important reason seems to be our everyday-life expertise in identifying people by their faces. It was found that people can be trained to recognize individual non-face objects, and thereby develop analysis patterns that are similar to those used in face recognition (e.g., Diamond & Carey, 1986; Gauthier, Williams, Tarr & Tanaka, 1998).

Other researchers presented participants with face images filtered by different spatial frequencies and found that only a rather narrow band of spatial frequencies (about 6 to 12 cycles per face width) contributes significantly to the recognition of a face (e.g., Näsänen, 1999). Again, this finding does not apply to non-face objects, even if individual objects of the same class are to be distinguished (Biederman & Kolacsai, 1997).

Moreover, face recognition has received considerable attention in machine vision research (e.g., Phillips, Moon, Rizvi & Rauss, 2000; Senior, Hsu, Mottaleb & Jain, 2002; Zhou., Krueger & Chellappa, 2003). Despite these immense efforts, however, even the currently best vision algorithms achieve face recognition rates that are far below the ones of a human observer.

We believe that a better understanding of the mechanisms underlying human face recognition will be beneficial to both the fields of medicine and machine vision. In the present study, we applied the sophisticated method of gaze-contingent windows to a psychophysical eye-movement study of a face recognition task in order to broaden our understanding of the underlying perceptual and attentional processes. The gaze-contingent window technique provides powerful experimental control and has been used extensively in reading, scene perception, and more recently in visual search studies (e.g. Bertera & Rayner, 2000; McConkie & Rayner, 1975; Pomplun, Reingold & Shen, 2001; Saida & Ikeda, 1979; see Rayner, 1998, for a review).

In most of its applications, this technique obscures all objects from view except those within a certain window that is continually centered on the participant's current gaze position. The window position changes across fixations to follow the gaze position. For example, in a study by McConkie and Rayner (1975), participants read text that was masked outside a visual window that included the fixated character and a number of characters to the left and to the right. Only the text within the window was legible. The visual span in reading was assessed by varying the window size across trials and determining the smallest window size that allowed participants to read with normal speed.

In the present study, participants were presented with images of famous and non-famous faces and had to indicate whether they recognized the displayed person or not. While viewing the images, a gaze-contingent window was administered with its size varying across trials. This allowed us to address the following questions: First, to what extent does face recognition rely on the simultaneous availability of the entire face features? Second, from which positions in the image and in what manner do participants acquire information about a face when their peripheral vision is restricted? It is well known that saccades during unrestricted face viewing tend to be aimed at the region formed by the eyes, nose, and mouth (e.g., Yarbus, 1967). However, where and how do participants gather information if they have to do it sequentially and be as efficient as possible? Third, how can the moment of recognition be characterized? Is it possible to determine this moment based on psychophysical data?

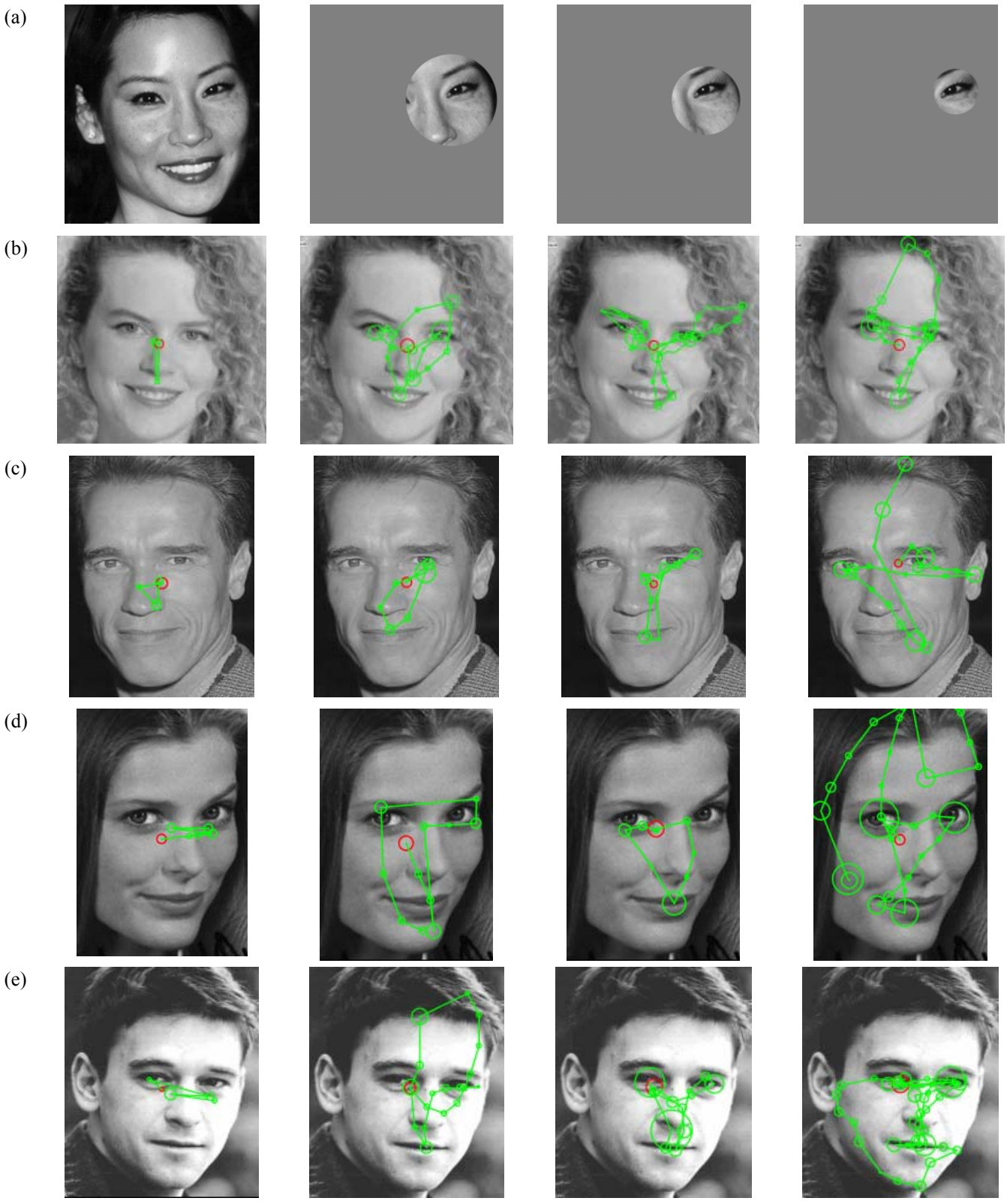

Figure 1: Sample stimuli used in the present study – each column represents one of the viewing conditions. From left to right column: unrestricted, large window, medium window, and small window. (a) Illustration of the different window sizes. (b) - (e) Sample gaze trajectories for each of the four stimulus categories (from top to bottom row): famous females, famous males, non-famous females, and non-famous males. Fixations are shown as circles with their size indicating fixation duration; the initial fixation is displayed in red color.

Method

Participants. Twenty students of the University of Massachusetts at Boston (ten females and ten males) participated in the present study. All of them had normal or corrected-to-normal vision. They were naïve with respect to the purpose of the study and were paid $10 for their participation.

Materials. We prepared 80 face images to serve as stimuli – 20 in each of the following four categories: famous females, famous males, non-famous females, and non-famous males. We chose the most popular American actresses and actors for the "famous" categories, while foreign actresses and actors, who had never appeared in international movies, were chosen for the "non-famous" categories. These grayscale images subtended an area of about 18° horizontally and 24° vertically on the screen of a 21-inch monitor. In the gaze-contingent window trials, the display area outside a circular, gaze-centered window was replaced with plain gray color. Four different viewing conditions were included in the experiment: unrestricted, large window (diameter of 8.2°), medium window (diameter of 5.5°), and small window (4.1°). These window sizes are illustrated in Figure 1a.

Apparatus. Eye movements were measured with an SR Research Ltd. EyeLink-II system. After a calibration procedure that was typically completed in less than a minute, gaze-position error was below or equal to 0.5 degrees of visual angle. The temporal resolution of the system was 2 ms. The gaze-contingent window followed the participant's gaze position with an average delay of 12 ms.

Procedure. Prior to each trial, participants were asked to fixate a marker in the center of the display. Following a button press, a face display was presented. As soon as participants had decided whether the depicted person was famous or non-famous, they terminated the trial by pressing one out of two buttons indicating their decision. Each participant was presented with each of the 80 stimuli exactly once, resulting in 80 trials per participant. The trials were administered in eight blocks of ten successive trials. Each of the four viewing conditions was applied in two of these blocks. The order of blocks and stimuli as well as the combination of stimuli with viewing conditions was systematically varied across participants.

Results and Discussion

Figures 1b to 1e show sample gaze trajectories for different stimuli across the four viewing conditions. Notice that the four trajectories for the same stimulus were generated by different participants, because each participant saw each stimulus only once. Two things can clearly be observed: First, in the unrestricted viewing condition, only a few central fixations were performed; the parafoveal and peripheral information of most of the face seems to be sufficient for successful face recognition. Second, when the gaze-contingent window was implemented, participants produced more fixations and directed them also at features that would normally not require foveal inspection, such as the hair or the ears, but obviously hold important information for the face recognition process. This effect of the gaze-contingent window on the eye-movement patterns generally increased with decreasing window size.

The quantitative analysis of the empirical data included the "standard" variables response time, proportion of correct responses, fixation duration, and saccade amplitude, but also the variables area coverage per trial and relative pupil size (see below). Interestingly, four-way analyses of variance (ANOVAs) for each of these variables (factors: viewing condition, stimulus recognizability, stimulus gender, and participant gender) revealed no significant effect by the factors stimulus gender or participant gender or their interaction. In other words, the gender of the participants or the people shown in the stimuli had no significant influence on any of the obtained variables. Therefore, in the following analyses, data were collapsed over these factors and only two-way ANOVAs (factors: viewing condition and stimulus recognizability) were conducted.

Response time was found to significantly depend on the viewing condition, $F(3; 57) = 80.93$, $p < 0.001$ as well as on recognizability (famous faces vs. non-famous faces), $F(1; 19) = 8.80$, $p < 0.01$. The interaction between the two factors also reached significance, $F(3; 57) = 2.77$, $p = 0.05$. As can be seen in Figure 2a, response time increased with smaller window size for both famous faces (no window: 1.90 s; large window: 4.57 s; medium window: 6.57 s; small window: 9.34 s) and non-famous faces (1.82 s, 5.69 s, 7.60 s, and 11.33 s). With more severe viewing restriction, response time became increasingly longer for non-famous faces as compared to famous ones. The pattern of these findings was expected, because restricting the participants' parafoveal and peripheral vision obviously makes their task more difficult. Detecting a familiar face should on average be faster than deciding that a face is unfamiliar, because before a negative decision can be made, all reasonable possibilities for a match have to be considered. With a smaller field of view, this effect increases, because more information needs to be obtained for a negative decision. The only unexpected finding is the pure magnitude of the response time difference imposed by the window manipulation.

The *proportion of correct responses* was also significantly influenced by the viewing condition, $F(3; 57) = 21.92$, $p < 0.001$, while there was neither an effect by recognizability, $F < 1$, or an interaction between the factors, $F < 1$. As shown in Figure 2b, there was no tradeoff between participants' response time and accuracy, but actually the opposite effect was found: The proportion of correct responses strongly decreased with more severe viewing restriction (91.3%, 78.0%, 67.8%, and 68.7%). Given that a participant who just gives random responses would reach an average of 50% correct responses, this result indicates a dramatic decrease in performance accuracy. The fact that there was no significant difference between famous and non-famous faces demonstrates that participants were not biased towards giving positive responses or towards giving negative responses. The latter would have been found if participants had just clicked the "non-famous" button

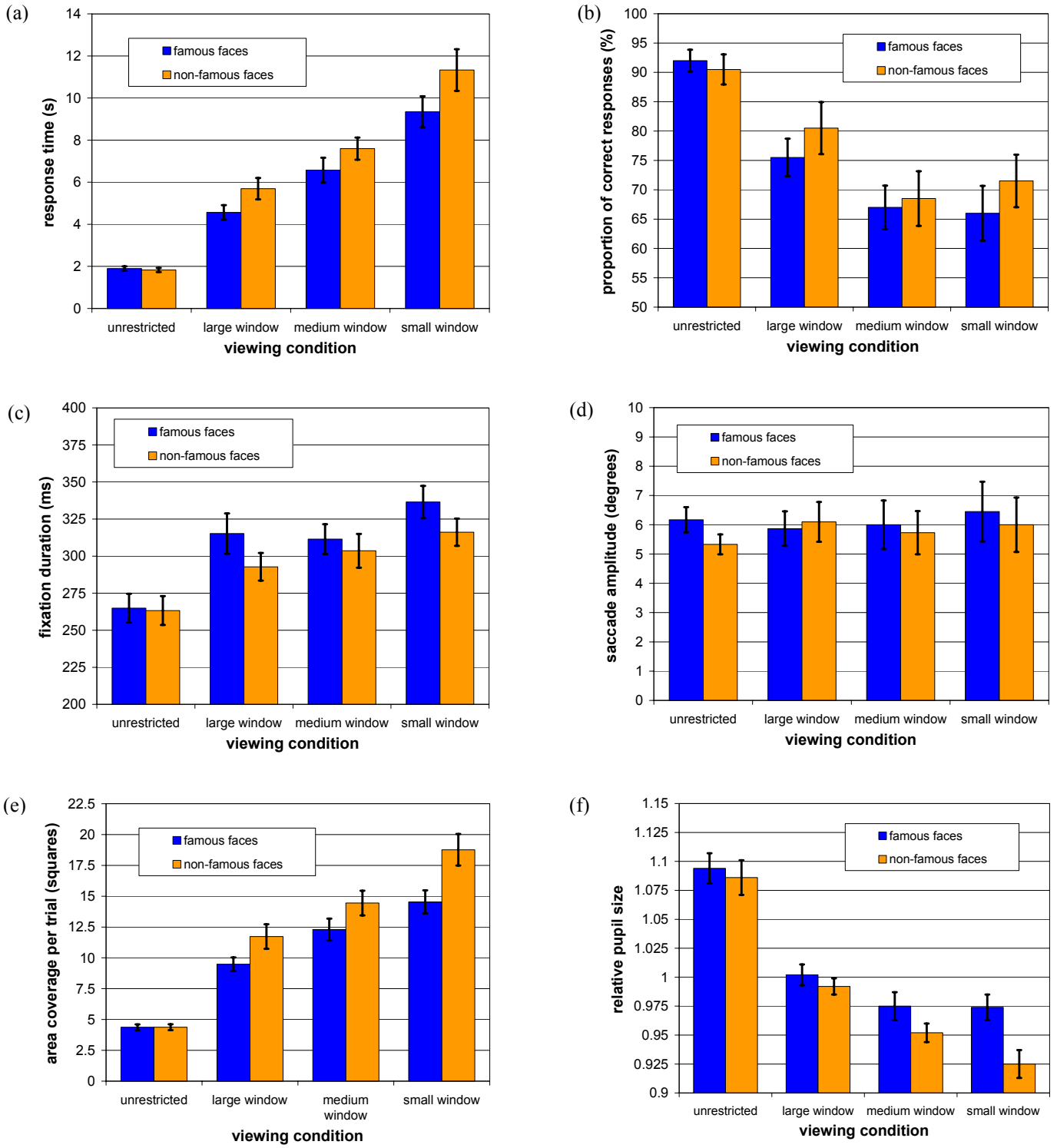

Figure 2: Psychophysical measurements obtained in the present study: (a) response time, (b) proportion of correct answers, (c) fixation duration, (d) saccade amplitude, (e) area coverage per trial, and (f) relative pupil size. Notice that the depicted interval of the variable values does not always start at 0.

whenever they did not recognize a face immediately, instead of making an effort to verify their first impression. Therefore, the analysis of the proportion of correct responses provides evidence for the participants of the present study to perform their task according to the instructions.

A variable that is analyzed in almost all eye-movement experiments is *fixation duration*. The duration of a fixation indicates how long the local information in a display was processed, which includes the duration necessary to program the subsequent saccade. In the present experiment, we found a significant effect by the viewing condition on fixation duration, $F(3; 57) = 23.15$, $p < 0.001$. The factor recognizability also exerted a significant effect, $F(1; 19) = 7.84$, $p < 0.05$, while there was no interaction between the two factors, $F(3; 57) = 1.33$, $p > 0.2$. In Figure 2c, fixation duration can be seen to increase with smaller windows for both famous faces (266 ms, 315 ms, 312 ms, and 336 ms) and non-famous faces (263 ms, 293 ms, 304 ms, 316 ms). Since smaller gaze-contingent windows reduce the amount of information that is available near the fixation point, there are two likely factors that determine this pattern of results: First, the smaller the window, the more effort is required to merge the available visual information with the current representation of the face in visual working memory. Second, for efficient task performance, a smaller window increases the necessity to aim saccades at locations where the most useful information is assumed to be located; consequently, the programming of these saccades requires more time. The finding of longer fixations for famous than for non-famous faces (see Figure 2c) is more puzzling; one possible explanation is that the recognition process itself causes one or more prolonged fixations – such fixations do not occur in non-famous faces. We conducted another analysis, reported later in this section, to test this hypothesis.

Another variable is routinely analyzed in eye-movement studies, namely *saccade amplitude*, measuring the length of saccades in degrees of visual angle. Short saccades can indicate fine-grained processing of local information, whereas long saccades often signify low information content or superficial scanning of local visual input. In the present context, we might expect saccades to become shorter with decreasing window size in order to uncover contiguous patches of an image. However, no influence by the viewing condition on saccade amplitude was found, $F < 1$, and recognizability showed no effect either, $F(1; 19) = 2.83$, $p > 0.1$. There was no interaction between the factors, $F(3; 57) = 1.74$, $p > 0.1$. As shown in Figure 2d, saccade amplitude (famous faces: 6.17°, 5.87°, 6.00°, and 6.45°; non-famous faces: 5.33°, 6.10°, 5.73°, and 6.00°) is unaffected by the recognizability of faces or the viewing condition. This result suggests that saccadic endpoints are not chosen to completely inspect local areas of the image by patching together adjacent pieces of visual information. Instead, saccades are aimed at positions that are assumed to contain the most significant information for the famous versus non-famous decision. This interpretation is in line with the view that fixations become longer with decreasing window size because of increased effort in the programming of saccades.

If it is true that the additional saccades induced by smaller gaze-contingent windows are aimed at "foraging" for useful information wherever in the image it is suspected, rather than inspect focused areas more thoroughly, then this behavior should be quantitatively reflected in the eye-movement data. In order to investigate this, we analyzed the eye-movement variable *area coverage per trial*. To compute this variable, we divided the stimulus area into 9 (horizontally) by 12 (vertically) squares. For each trial, we calculated the area coverage as the number of different squares that contained at least one fixation. This number should increase with the amount of "information foraging" (as opposed to focused examination) performed by participants. We found the area coverage per trial to be significantly influenced by the viewing condition, $F(3; 57) = 96.42$, $p < 0.001$, and by recognizability, $F(1; 19) = 13.50$, $p < 0.01$. The interaction between these factors was also significant, $F(3; 57) = 5.78$, $p < 0.01$. As shown in Figure 2e, the pattern of results for area coverage per trial is very similar to the one for response times (Figure 2a): Area coverage per trial increases strongly with decreasing window size, both for famous faces (4.37, 9.49, 12.30, and 14.54 squares) and for non-famous ones (4.38, 11.73, 14.45, and 18.77 squares). Area coverage is also smaller for famous faces than for non-famous ones, with this difference being more pronounced for smaller windows. This finding supports our assumption that peripheral restriction of information induces an exploration strategy that guides saccades towards the most promising new locations in the stimulus.

Finally, a variable that usually receives less attention, although it is a "by-product" of video-based eye tracking, is *pupil size*. Pupil size is known to depend on factors such as the luminance in the visual field or the cognitive activation of a person (e.g. Kahneman, 1973; Pomplun & Sunkara, 2003). Since we were only interested in relative changes in pupil size, we divided all measurements by the participants' initial pupil size after the eye tracker setup. Consequently, values above 1 indicate a dilated pupil, while values below 1 signify a contracted pupil. We found a significant influence of the viewing condition on pupil size, $F(3; 57) = 42.58$, $p < 0.001$, and also a significant influence by recognizability, $F(1; 19) = 8.37$, $p < 0.01$. There was no interaction, $F(3; 57) = 1.81$, $p > 0.1$. Figure 2f illustrates that pupil size clearly decreases with smaller window size for both famous faces (1.094, 1.002, 0.975, and 0.974) and non-famous faces (1.086, 0.992, 0.952, 0.925). This could be explained by the smaller amount of information that is available for processing at any given time during the trial. However, why was the pupil larger for famous than for non-famous faces?
A possible explanation is that the *moment of recognizing* a famous face has an impact on the eye-movement data. In an earlier, unpublished study on a face search experiment, we observed that the moment of recognizing the presence of a face tended to coincide with a prolonged fixation and temporarily dilated pupils. If such a reaction also occurs at the moment of recognizing a known face, then in the present study this would not only explain the greater in pupil size, but also the longer fixations measured for famous as compared to non-famous faces. To test this hypothesis, we

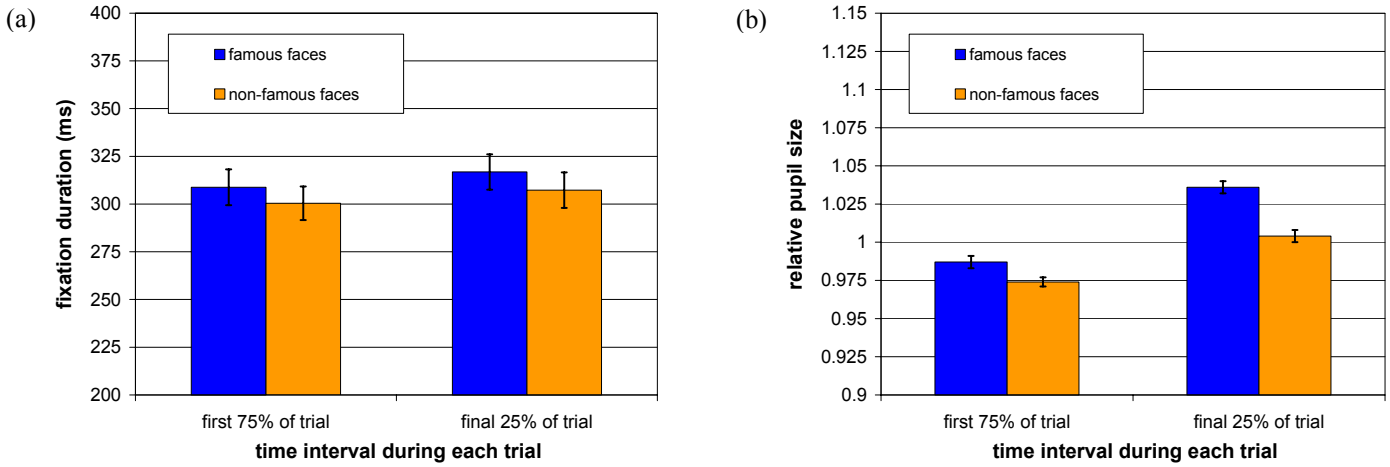

Figure 3: (a) Fixation duration and (b) pupil size analyzed separately during the first 75% and the last 25% of each trial.

separated the data for the first 75% of the duration of each trial from the final 25%. The moment of recognition is most likely to occur during the last 25% of a trial, so if we found the recognizability effects on fixation duration and pupil size to only occur during the last phase, as indicated by an interaction of the factors recognizability and time interval, it would support the moment of recognition interpretation.

For each of the two variables, we therefore conducted a two-way (recognizability and time interval) ANOVA. While there was no significant interaction for fixation duration (Figure 3a), $F < 1$, it was found for pupil size, $F(1; 19) = 6.92$, $p < 0.05$. Figure 3b shows that the difference in pupil size between famous and non-famous faces clearly increases during the final 25% of the trials. This finding suggests that the moment of recognizing a face may be associated with pupil dilation.

All in all, the present study has shown that the simultaneous availability of the entire face information is crucial for efficient face recognition, which supports the view that face recognition is a holistic process that heavily relies on parafoveal and peripheral input. Restricting this input has provided us with insight into which essential information was eliminated and needed to be foveally processed instead. Finally, we have found that the moment of recognizing a face may be indicated by a dilated pupil. This line of research is only at its beginning, and we hope to inspire other researchers to consider the technique of gaze-contingent windows for their face recognition and other perceptual studies.

References

Bertera, J.H. & Rayner, K. (2000). Eye movements and the span of the effective visual stimulus in visual search. *Perception & Psychophysics*, 62, 576-585.

Biederman, I. & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London*, 352, 1203-1219.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107-117.

Gauthier, I., Williams, P., Tarr, M.J. & Tanaka, J. (1998). Training "greeble" experts: A framework for studying expert object recognition processes. *Vision Research*, 38, 2401-2428.

Kahneman, D. (1973). *Attention and effort*. New Jersey: Prentice Hall.

Maurer, D., LeGrand, R. & Mondloch, C.J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, 6, 255-260.

McConkie, G.W. & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17, 578-586.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, 39, 3824-3833.

Phillips, P.J., Moon, H., Rizvi, S.A. & Rauss, P.J. (2000). The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 1090-1104.

Pomplun, M., Reingold, E.M. & Shen, J. (2001). Peripheral and parafoveal cueing and masking effects on saccadic selectivity in a gaze-contingent window paradigm. *Vision Research*, 41, 2757 - 2769.

Pomplun, M. & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In D. Harris, V. Duffy, M. Smith & C. Stephanidis (Eds.), *Human-Centred Computing: Cognitive, Social, and Ergonomic Aspects.* Vol. 3 of the Proceedings of the 10th International Conference on Human-Computer Interaction, HCI 2003, Crete, Greece.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Saida, S. & Ikeda, M. (1979). Useful visual field size for pattern perception. *Perception & Psychophysics*, 25, 119–125.

Senior, A., Hsu, R., Mottaleb, M.A. & Jain, A.K. (2002). Face detection in color images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 696-706.

Yarbus, A.L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zhou, S., Krueger, V. & Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91, 214-245.